DOCTOR OF SCIENCES

Méta-analyse de damiers à ADN pour l'identification de gènes impliqués dans l'hypoxie et causant le phénotype métastatique : mesure de leur expression dans des cellules de différents potentiels métastatiques en hypoxie et en normoxie

Pierre, Michael

Award date:
2011

Awarding institution:
University of Namur

META-ANALYSE DE PUCES A ADN POUR L'IDENTIFICATION DE GENES IMPLIQUES DANS L'HYPOXIE ET PROVOQUANT LE PHENOTYPE METASTATIQUE

MESURE DE LEUR EXPRESSION DANS DES CELLULES DE DIFFERENTS POTENTIELS METASTATIQUES EN HYPOXIE ET NORMOXIE

Directeurs :
Eric DEPIEREUX
Carine MICHIELS
Rapporteurs :
Xavier DE BOLLE
Thierry ARNOULD
Thierry COCHE

Thèse de doctorat présentée par
Michael PIERRE
en vue de l'obtention du diplôme de
docteur en science biologique
Options bioinformatique, biostatistique
et biologie cellulaire
Année académique 2011/2012

Remerciements

Voici quatre ans, commencait la thèse dont ce manuscrit est l'objet. Alors que les perspectives que laissait entrevoir mon passage devant le jury du FRIA étaient plutôt mauvaises, je me lancais dans la rédaction d'un nouveau projet pour le Télévie dont la thématique biologique était plus précise. Le FRIA finalement obtenu de justesse, je restai avec ce projet qui, malgré ce départ quelque peu difficile, connut bien plus de hauts que de bas. Et ceci n'est pas un hasard ! Au-delà de ces pages, de nombreuses personnes ont contribué, de près ou de loin, à ce que ce travail soit au final une belle réussite. C'est donc bien normalement que je voudrais adresser mes remerciements à quelques personnes qui ont rendu tout ceci possible.

D'abord, je voudrais remercier mon promoteur Eric Depiereux. Cela fait maintenant plus de quatre ans que je travaille dans son équipe et cela a toujours été avec un grand plaisir. Depuis le début de mon mémoire, il a toujours été là pour me guider et m'encourager. Il m'a toujours laissé faire mes choix et les a toujours soutenus ; sans lui je n'en serais pas là où j'en suis aujourd'hui.

Ensuite, je dois également un grand merci à ma co-promotrice Carine Michiels. Alors qu'elle me connaissait à peine, elle a accepté de parier sur moi et a soutenu le projet depuis le premier jour. Elle n'a jamais hésité à sacrifier son temps et son énergie pour m'apprendre les techniques de laboratoire, corriger mon travail ou m'encourager pour faire de ce travail ce qu'il est.

Je voudrais aussi remercier les membres de mon jury : Thierry Arnould, Xavier De Bolle et Thierry Coche. Le temps et l'intérêt que vous avez porté à ce travail me touche énormément et grâce à vous ce manuscrit est devenu bien meilleur à tous les niveaux.

Je remercie également le FRIA pour avoir financé ce projet et son jury pour avoir cru en moi et m'avoir laissé la chance de modifier mon projet. Sans cela rien de ce qui se trouve dans ce manuscrit n'aurait été possible.

Mes remerciements vont aussi aux personnes qui m'ont entouré pendant plus de quatre ans dans l'équipe de bioinfo : Anthoula Gaigneaux, Fabrice Berger, Benoît De Hertogh, Eric Bareke, Bertrand De Meulder, Sophie Depiereux et Raphaël Helaers. Non seulement c'était un plaisir de vous avoir comme collègues, mais en plus je suis ravi de vous compter parmi

mes amis. Les concerts et les colloques que nous avons faits ensemble resteront gravés dans ma mémoire !

Merci à toutes les personnes de l'URBM. Je ne vais pas faire un listing du personnel, mais je voudrais adresser mes remerciements à Jean-Jacques Letesson, Jean-Yves Matroule, Damien Hermand et Isabelle Housen pour l'intérêt qu'ils ont porté à mon travail lors des réunions en blue room ; à Etienne, Françoise, Matthieu, Charles, Marie, Marie-Alice, Michaël, Sophie et les autres pour ces quatre années passées en URBM. Évidemment un grand merci à ceux avec qui j'ai passé des moments si amusants hors du labo : Aurore, Mira, Fanélie, Delphine, Marie, Max, Anne-Michèle et Caro.

Merci aussi à toutes les personnes de l'URBC pour m'avoir accueilli dans votre unité. De nouveau, pas de listing, mais des remerciements particuliers à Annick, Kayleen, Lionel Leclere, Lionel Flamant, Hélène, Marie, Manu, Guillaume, Aurélie, Aurélia, Antoine, Martine, Guy et les autres pour leur aide ô combien précieuse pour un pauvre bioinformaticien et surtout pour leur amitié.

Je voudrais aussi remercier Yves Poumay et Conny Mathay pour m'avoir donné l'opportunité de travailler sur un de leurs projets avec eux. En plus d'avoir été instructif et passionnant, j'ai rencontré deux personnes que j'apprécie particulièrement que ce soit sur le plan professionnel ou humain.

Je remercie également tous les professeurs des facultés de Namur qui m'ont formé. Sans vous je ne serais pas allé aussi loin.

Je remercie mes parents pour avoir cru en moi, m'avoir encouragé et m'avoir soutenu depuis toujours. Merci de m'avoir laissé ma chance, de m'avoir laissé faire mes erreurs et de m'avoir redonné ma chance. Je suis bien conscient qu'il n'a pas dû être toujours facile de me porter jusque là. Sans vous, tout ça n'aurait pas été possible et je ne serais pas celui que je suis, c'est pour cela que je vous remercie.

Enfin, je voudrais remercier une dernière personne. Il s'agit de celle qui me soutient et me supporte (et ça ne doit pas toujours être facile) tous les jours de ma vie. C'est elle qui rend ma vie plus belle et pour qui je veux aller toujours plus loin. Sans toi, ma vie serait bien fade. Amandine, je te remercie pour tout ce que tu es.

Encore mille merci à tous et à bientôt en d'autres temps et d'autres lieux !

TABLE DES MATIERES

TABLE DES ABREVIATIONS

^{32}P	Phosphore 32
ADN	Acide désoxyribonucléique
ADNc	ADN complémentaire
AE	ArrayExpress
AKT/PKB	v-akt murine thymoma viral oncogene homolog 1
AMF	Autocrine Motility Factor
ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
Apaf-1	Apoptotic peptidase activating factor 1
ARN	Acide ribonucléique
ARNm	ARN messager
Bak	BCL2-antagonist/killer 1
Bax	BCL2-associated X protein
Bcl-2	B-cell CLL/lymphoma 2
Bcl-w	BCL2-like 2
Bcl-x_L	BCL2-like 1
CDF	Chip Definition File
CDH1	Cadherin 1, type 1, E-cadherin (epithelial)
CDK4	Cyclin-dependent kinase 4
CFL2	Cofilin 2 (muscle)
ChIP on chip	Chromatin ImmunoPrecipitation on chip
COX-2	Prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)
DAVID	Database for Annotation, Visualization and Integrated Discovery
DDC	Diethyl 1,4-dehydro-2,3,6-trimethyl-3,5-pyridine decarboxylate
DDR	Double Data Rate
E2F	E2F transcription factor
E7	E7 protein
E-cadhérine	Epithelial cadherin
EGF-R	Epidermal Growth Factor – Receptor
FADD	Fas (TNFRSF6)-associated via death domain
FAERI	Functional analysis: evaluation of response intensities
FAS	Fas (TNF receptor superfamily, member 6)
FDR	False Discovery Rate
FGF	Fibroblast Growth Factor
FWER	Family-Wise Error Rate
gb	Gigabyte
GCRMA	GeneChip Robust Multi-array Analysis
GEMS	Gene Expression MetaSignatures
GEO	Gene Expression Omnibus
Ghz	Gigahertz
GLUT1	Solute carrier family 2 (facilitated glucose transporter), member 1
GSE	GEO Series
HER2/neu	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)
HG	Human Genome
HIF-1	Hypoxia-Inducible Factor-1

HRE	Hypoxia Response Element
IGF-1/2	Insulin-like growth factor 1/2
IGF-1R	Insulin-like growth factor 1 receptor
IL-3	Interleukin 3 (colony-stimulating factor, multiple)
IL-3R	Interleukin 3 receptor, alpha (low affinity)
iNOS	Isoform Nitric oxide synthases
IRF3	Interferon regulatory factor 3
KEGG	Kyoto Encyclopedia of Genes and Genomes
LIMK1	LIM domain kinase 1
LS	Latin Square
Mad	MAX dimerization protein 1
MAPK	Mitogen-activated protein kinase
MAQC	MicroArray Quality Control
MAS	MicroArray Suite
Max	MYC associated factor X
MCF-7	Michigan Cancer Foundation – 7
Mcl-1	Myeloid cell leukemia sequence 1 (BCL2-related)
MDA-MB-231	M.D. Anderson - metastatic breast – 231
MEC	Matrice extracellulaire
MIAME	Minimum Information About a Microarray Experiment
miRNA	MicroARN
MM	MisMatch
MMP	Matrix metalloproteinase
MPT	Mitochondrial permeability transition
MTGDR	Meta Threshold Gradient Descent Regularization
MTT	3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide
MYC	v-myc myelocytomatosis viral oncogene homolog (avian)
N-cadhérine	Neural cadherin
NCBI	National Center for Biotechnology Information
NF2	Neurofibromin 2 (merlin)
NFκB	Nuclear Factor kappa B
Noxa	Phorbol-12-myristate-13-acetate-induced protein 1
p15INK4B	Cyclin-dependent kinase inhibitor 2B (p15, inhibits CDK4)
p38	Mitogen-activated protein kinase 1
p53	Tumor protein p53
PAK1	p21 protein (Cdc42/Rac)-activated kinase 1
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PDGF	Platlet-Derived Growth Factor
PHD	Prolyl-4-hydroxylase
PI3K	Phosphatidylinositol 3'-kinase
PM	Perfect Match
polyA	Poly adénosine
pRb	Retinoblastoma protein
PTEN	Phosphatase and tensin homolog
Puma	BCL2 binding component 3
PVDF	Polyvinylidene fluoride

RAC2	ras-related C3 botulinum toxin substrate 2 (rho family, small GTP binding protein Rac2)
RAF	Zinc fingers and homeoboxes 2
RAID	Redundant Arrays of Inexpensive Disks
RAS	Ras protein
RMA	Robust Multi-array Analysis
ROC	Receiver Operating Characteristic
RPMI	Roswell Park Memorial Institute
RT-PCR	Reverse Transcription PCR
SDS	Sodium Dodecyl Sulfate
siRNA	Small interfering RNA
SLUG	Snail homolog 2 (Drosophila)
SMAD4	SMAD family member 4
SNAIL	Snail homolog 1 (Drosophila)
SNP	Single Nucleotide Polymorphism
tBid	BH3 interacting domain death agonist
TBK1	TANK-binding kinase 1
TCF3	Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)
TGFα	Tumor Growth Factor α
TGFβ	Tumor Growth Factor β
TICAM1	Toll-like receptor adaptor molecule 1
TLR	Toll-like recepteur
TNF-R1	Tumor Necrosis Factor Receptor superfamily, member 1A
TPF	True Positive Fraction
TRAF3	TNF receptor-associated factor 3
U	Ubiquitine
uPA	Urokinase-type Plasminogen Activator
uPAR	Urokinase-type Plasminogen Activator Receptor
VAV2	Vav 2 guanine nucleotide exchange factor
VEGF	Vascular Endothelial Growth Factor
VHL	von Hippel-Lindau
ZFHX1A	Zinc finger E-box binding homeobox 1
ZFHX1B	Zinc finger E-box binding homeobox 2

I. INTRODUCTION

1. Avant-propos

Ce travail repose sur deux grands axes, l'un biologique, l'autre méthodologique. Bien que ce soit la problématique biologique qui ait motivé toutes les approches qui seront présentées dans le présent manuscrit, des résultats méthodologiques ont également été obtenus. C'est pourquoi cette introduction sera divisée en deux grandes parties, dans l'objectif de présenter au lecteur l'importance de chacun des axes.

La première partie abordera donc la problématique du cancer et des métastases. Elle tentera d'éclairer le lecteur quant au développement de cette maladie d'un point de vue moléculaire. Elle expliquera également le rôle que joue l'hypoxie dans l'apparition des métastases. Dans la seconde partie, les puces à ADN seront décrites tant du point de vue de leur conception que de l'analyse des résultats qu'elles génèrent.

L'objectif de ce travail est de tirer parti de l'importante quantité de jeux de données obtenus à partir de puces à ADN et étudiant l'expression génique différentielle dans les métastases et/ou en réponse à l'hypoxie ; ceci afin de générer de nouvelles hypothèses quant au développement du phénotype métastatique des cellules cancéreuses. Pour réaliser cet objectif une méthode d'analyse originale a été développée. Dans un second temps, les hypothèses générées par cette méthode ont été validées sur cellules.

2. La problématique biologique

2.1. Le cancer

2.1.1. Généralités

Le cancer est une maladie qui est le résultat de la transformation d'une cellule dont la prolifération devient incontrôlée. S'en suivent un envahissement du tissu et une éventuelle dissémination dans d'autres parties de l'organisme. Puisque n'importe quelle cellule peut devenir transformée, il y a autant de types de cancers qu'il y a de types de cellules [1]. Généralement, le type de cancer se définit par l'organe qu'il touche : cancer du poumon, cancer du pancréas, cancer du cerveau, …

Les facteurs causant le cancer sont également de plusieurs types. Ceux-ci peuvent être endogènes ou exogènes.

Enfin, bien que le cancer soit une maladie que l'homme a identifiée très tôt dans son histoire et que la médecine actuelle ait fait des progrès considérables, elle reste une pathologie causant encore beaucoup de décès. Tous les mécanismes provoquant le développement et la croissance des tumeurs ne sont pas encore connus et compris. Toute avancée ouvre donc la porte à de nouvelles opportunités de combattre cette maladie.

2.1.2. Développement de la maladie

Pour sortir de la phase G_0, que l'on considère comme ne faisant pas partie du cycle cellulaire car les cellules sont dans un état quiescent [2], et pour progresser dans le cycle cellulaire, pendant lequel les cellules préparent et effectuent la division (Figure 1), toute une série de protéines sont requises ; la transcription de leur gène et la traduction des transcrits sont donc augmentées spécifiquement [3]. Une ou plusieurs mutations dans l'un des gènes codant pour ces protéines peut être à l'origine d'un cancer engendrant une prolifération cellulaire incontrôlée et provoquant le développement d'une tumeur [4]. Il faut aussi signaler que la mutation d'une seule des copies d'un oncogène peut être suffisante pour que la cellule acquière la capacité de se diviser plus rapidement.

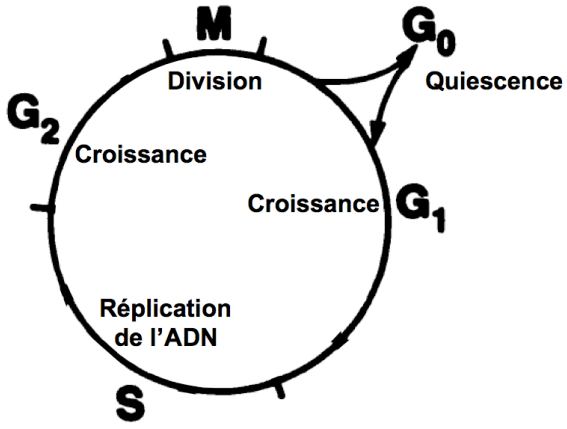


Figure 1. *Représentation schématique du cycle cellulaire. Durant la phase G_0, les cellules ne prolifèrent pas. Avant de se diviser, elles doivent passer par trois phases. Lors de la phase G_1, qui succède à la phase G_0 ou à la mitose précédente, les cellules synthétisent de nombreuses enzymes, en particulier celles nécessaires à la réplication de l'ADN. Si les cellules ne présentent pas de dommages à l'ADN, elles passent le point de contrôle et entrent dans la phase S où l'ADN est répliqué. Vient ensuite la phase G_2 durant laquelle les cellules grossissent. Enfin, si les cellules passent le nouveau point de contrôle, elles effectuent la mitose (source : figure modifiée de [5]).*

D'autres protéines ont pour rôle d'empêcher la division cellulaire en provoquant l'apoptose ou en bloquant le cycle cellulaire. Elles sont codées par des gènes suppresseurs de tumeur. Une mutation au sein de l'un de ces gènes peut également être responsable du développement d'un cancer [6]. En effet, les gènes suppresseurs de tumeur servent de frein à la prolifération cellulaire. Si une cellule perd ce frein, rien ne l'empêche plus de se diviser. Il s'agit dans ce cas d'une mutation « perte de fonction », mutation récessive car les deux copies du gène suppresseur de tumeur doivent être mutées pour que la cellule perde la capacité d'enclencher l'apoptose ou de stopper son cycle.

Enfin, il faut noter qu'il existe des protéines chargées de réparer les mutations que subit l'ADN [7]. En effet, bien que le système de réplication de l'ADN soit très fiable, le taux d'erreurs n'est pas nul. De plus, les gènes codant pour les protéines de ces systèmes de réparation de l'ADN peuvent, eux aussi, subir des mutations. La perte de fonctionnalité de l'un de ces systèmes peut donc aussi être à l'origine du développement cancéreux [8].

En fait, la transformation d'une cellule normale en cellule cancéreuse est le résultat d'une série de mutations qui apparaissent selon une séquence particulière, qui peut être différente d'un cancer à l'autre. De plus, dans certains cas, l'apparition d'une mutation favorise l'apparition de la suivante. En effet, la mutation d'un oncogène, d'un gène suppresseur de tumeur ou d'un gène de réparation de l'ADN peut donner un avantage sélectif à la cellule si cette altération n'est pas réparée. Ainsi, l'accumulation de mutations, qui ne sont pas réparées et qui donnent un avantage sélectif, permet, à terme, à la cellule présentant les mutations de supplanter le reste de la population cellulaire. Ce processus s'appelle la transformation et s'étale souvent sur plusieurs années [9].

Un cancer commence donc par une première mutation qui n'est pas réparée et qui touche soit un oncogène, soit un gène suppresseur de tumeur, soit un gène de réparation de l'ADN. Un processus de sélection s'exerce et les mutations s'accumulent. À ce stade, il n'y a pas de conséquences fonctionnelles pour l'organisme. La cellule transformée prolifère de manière anarchique et un amas de cellules cancéreuses se forme. Si celles-ci envoient des signaux moléculaires aux cellules normales environnantes et que ces dernières y répondent, la tumeur naissante obtiendra les nutriments nécessaires à sa croissance. Si en grandissant la tumeur rompt la membrane basale, le cancer devient invasif. C'est à ce stade qu'une ou plusieurs cellules peuvent se détacher de la tumeur pour pénétrer dans un vaisseau sanguin ou dans un vaisseau lymphatique et se disséminer à d'autres parties de l'organisme. Ce phénomène s'appelle la métastase et sera détaillé plus loin.

Outre l'instabilité génomique [10], qui vient d'être détaillée, il existe un autre processus qui permet le développement cancéreux : l'inflammation. Pendant longtemps, les pathologistes ont observé que certaines tumeurs contenaient un grand nombre de cellules du système immunitaire, tant inné qu'adaptatif [11]. Avec la découverte de nouveaux marqueurs pour identifier les différents types de cellules du système immunitaire, on sait maintenant que toutes les tumeurs comportent de telles cellules, même en faible nombre [12]. Bien qu'il y ait de plus en plus d'arguments qui montrent que la réponse immunitaire a pour objectif d'éliminer la tumeur, elle a aussi, paradoxalement, pour effet de promouvoir le développement de la tumeur. En effet, les cellules immunitaires peuvent fournir aux cellules myéloïdes suppressives se trouvant dans la tumeur des facteurs induisant la suppression immunitaire. Ces cellules favorisent également l'angiogenèse et l'invasion [13]. De plus, il faut noter que les cellules inflammatoires sont capables de produire des molécules, comme les dérivés réactifs de l'oxygène, qui augmentent l'instabilité génomique des cellules cancéreuses, accélérant ainsi la transformation [14].

2.1.3. Types de cancers

Dans cette section, nous allons présenter une classification tout à fait générale. Chaque catégorie comprend, en effet, différents niveaux correspondant généralement à l'organe ou au type de cellule touché. Cependant, ce niveau de détail ne sera pas abordé ici.

- Les carcinomes. Ils touchent les épithéliums, c'est-à-dire les tissus uniquement composés de cellules [15].
- Les sarcomes. Ils touchent les tissus conjonctifs c'est-à-dire des tissus composés à la fois de cellules, mais aussi de fibres de collagène et d'autres éléments extracellulaires [16].
- Les lymphomes. Il s'agit d'une catégorie de cancers caractérisée par le fait que les tumeurs ne sont pas solides puisqu'elles sont composées de cellules hématopoïétiques ou cellules du sang [17].
- Les cancers des cellules germinales. Ce sont des tumeurs dérivant de la transformation et de la division anarchique des cellules totipotentes de l'organisme, c'est-à-dire les cellules qui sont capables de se spécialiser en n'importe quel type cellulaire [18].

- Les blastomes. Ce sont des tumeurs ressemblant aux tissus embryonnaires, ils sont, d'ailleurs, généralement trouvés chez les enfants [19].

2.1.4. Causes de la maladie

Cette section ne cherche en rien à établir une liste, qui ne serait de toute façon pas exhaustive, des différentes causes de cancer. Elle a seulement pour objet de distinguer deux grands types de facteurs en les illustrant par quelques exemples.

Les facteurs endogènes sont principalement des altérations génétiques et épigénétiques héréditaires, l'âge et des mutations provoquées par des causes endogènes. Les altérations génétiques héréditaires touchent les cellules germinales d'un individu qui les transmet dès lors à sa descendance. Cependant, la transformation est souvent la conséquence d'une interaction complexe entre l'altération génétique héritée et l'environnement. Un exemple d'altération épigénétique est la méthylation d'un gène suppresseur de tumeur le rendant ainsi silencieux [20]. L'âge est un autre facteur endogène pouvant causer le développement d'un cancer. En effet, comme vu précédemment, des mutations de gènes peuvent survenir sans que celles-ci ne soient réparées. Bien que cela puisse ne pas avoir de conséquences, avec l'âge, ces mutations non réparées s'accumulent et peuvent mener vers le cancer si elles touchent certains gènes particuliers décrits plus haut. Les espèces réactives de l'oxygène sont une autre cause endogène qui peut provoquer des dommages à l'ADN [21]. Une fois encore, si ces dommages ne sont pas réparés et qu'ils touchent des gènes sensibles, cela peut conduire au développement d'une tumeur.

Les facteurs exogènes comportent l'ensemble des agents qui peuvent provoquer le développement d'un cancer soit en attaquant directement l'ADN, comme certaines radiations dont les rayons ultraviolets du soleil [22], soit en rendant le milieu favorable à l'apparition de cellules cancéreuses, comme l'alcool qui provoque l'inflammation [23]. À titre d'exemples, on peut aussi citer le tabac [24], certains virus [25], l'alimentation et de nombreux agents chimiques comme facteurs exogènes pouvant provoquer le cancer.

2.1.5. Mécanismes cellulaires

Cette section va tenter d'aborder brièvement quelques-uns des grands mécanismes moléculaires observés dans les cellules au cours de leur transformation en cellules cancéreuses. Hanahan et Weinberg ont proposé que, malgré l'immense diversité de cancers et l'augmentation en complexité de l'information que la recherche accumule à propos de ces

maladies, il est possible de trouver des dénominateurs communs aux cancers. Ceux-ci sont huit capacités qu'acquièrent les cellules cancéreuses. Six capacités ont été mises en évidence dans un premier article en 2000 [26], ce sont l'autosuffisance en signaux de croissance, l'insensibilité aux signaux d'inhibition de croissance, l'échappement à l'apoptose, un potentiel de réplication illimité, la néoangiogenèse et la métastase (Figure 2). Ensuite, dans un second article paru début 2011 [27], Hanahan et Weinberg ont pointé deux capacités supplémentaires : la reprogrammation du métabolisme énergétique et l'échappement au système immunitaire. Selon Hanahan et Weinberg, ces huit capacités sont essentielles pour le développement d'une tumeur maligne et leur acquisition est due à l'instabilité du génome, qui a été évoquée plus haut, à l'inflammation, qui favorise le fonctionnement de ces huit capacités (Figure 2), et au microenvironnement des cellules cancéreuses. En effet, les tumeurs ne sont pas des masses isolées de cellules transformées, mais bien des tissus complexes formés, en plus des cellules cancéreuses, de cellules normales qui participent à l'acquisition et au développement des huit capacités. Ces huit capacités, qui sont acquises selon un ordre aléatoire, sont décrites ci-dessous.

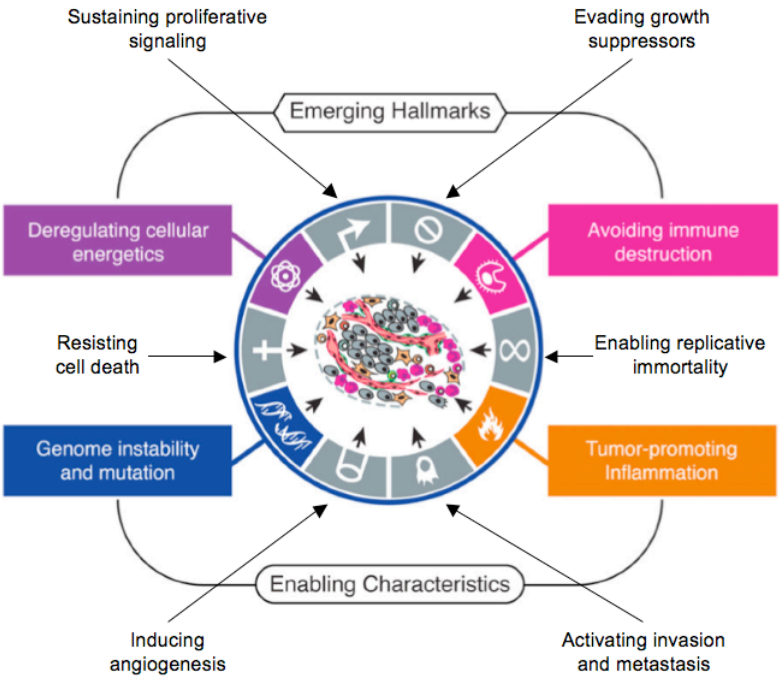


Figure 2. *La reprogrammation du métabolisme énergétique et l'échappement au système immunitaire sont les deux capacités qui ont été ajoutées aux six capacités pointées onze ans auparavant par Hanahan et Weinberg. L'instabilité du génome et l'inflammation sont deux causes proposées par les auteurs pour expliquer l'acquisition de ces huit capacités (source : figure modifiée de [27]).*

a) L'autosuffisance en signaux de croissance

Afin de passer de la phase G_0 à une division active, une cellule normale a besoin de recevoir des signaux. Ces signaux sont des molécules que perçoit la cellule via des récepteurs transmembranaires [28]. Pour initier la division cellulaire, la cellule doit percevoir plusieurs types de signaux : soit des facteurs de croissance se trouvant libres dans le milieu, soit des structures particulières de la matrice extracellulaire, soit des facteurs se trouvant à la surface des cellules voisines.

Dans le cas des cellules cancéreuses, une protéine sécrétée, résultant de la transcription et de la traduction d'un oncogène, peut mimer ces facteurs de croissance et aller se lier aux récepteurs transmembranaires qui transmettent le message d'initiation de la division cellulaire. Ce phénomène peut s'observer *in vitro*. En effet, une culture de cellules normales nécessite l'apport de facteurs de croissance extérieurs pour proliférer, alors qu'une culture de cellules cancéreuses ne demande pas cet apport car les cellules en sont devenues indépendantes en produisant leurs propres facteurs de croissance comme le PDGF (Platlet-Derived Growth Factor) ou le TGFα (Tumor Growth Factor α). Les cellules cancéreuses peuvent aussi envoyer des signaux pour stimuler les cellules normales de leur microenvironnement pour qu'elles leur fournissent des facteurs de croissance [29]. Cependant, des études récentes ont montré qu'un excès d'oncoprotéines comme RAS, MYC ou RAF pouvait mener une cellule vers un état de sénescence, où elle ne prolifère plus, ou vers l'apoptose [30]. Par exemple, des cellules en culture qui expriment fortement Ras peuvent entrer dans un état de sénescence, alors que les mêmes cellules exprimant moins Ras prolifèrent. À nouveau, les cellules cancéreuses s'adaptent à des hauts taux de facteurs de croissance car les voies de signalisation menant à la sénescence ou à l'apoptose sont inactives [31].

L'autosuffisance en facteurs de croissance peut résulter d'autres phénomènes que leur surexpression. Les récepteurs transmembranaires peuvent, eux aussi, subir des dérégulations [32], par exemple en étant surexprimés, comme l'EGF-R (Epidermal Growth Factor – Receptor) dans les cancers de l'estomac, du cerveau et du sein ou comme le récepteur à HER2/neu dans les carcinomes mammaires et de l'estomac, rendant, alors, la cellule hypersensible aux facteurs de croissance en y répondant plus rapidement [33]. Les récepteurs transmembranaires peuvent aussi être mutés de telle manière que la liaison à un facteur de croissance n'est plus nécessaire pour envoyer le message d'initiation de la division cellulaire, comme c'est le cas pour certaines versions tronquées de l'EGF-R [34].

Le dernier mécanisme par lequel une cellule cancéreuse peut devenir indépendante des facteurs de croissance est une dérégulation du système qui transmet le signal, perçu par les récepteurs transmembranaires, à l'intérieur de la cellule [35]. Cela implique qu'une (ou plusieurs) protéine(s) de ce système soi(en)t mutée(s). On observe, notamment, dans de nombreux cancers, une mutation de la protéine Ras, qui est un élément central dans la transmission du message d'initiation de la mitose. Ras devient alors indépendant des facteurs qui la régulent normalement dans la cascade de transmission. De plus, la version mutée de Ras peut interagir avec PI3K (phosphatidylinositol 3'-kinase) qui lui-même promeut la survie cellulaire. Il existe, normalement, des systèmes de contrôle qui empêchent des facteurs tels que PI3K d'avoir un effet prolongé. Cependant, on remarque dans les cellules cancéreuses que ces systèmes sont aussi mutés. C'est le cas, par exemple, de la phosphatase PTEN, qui a pour rôle de déphosphoryler le phosphatidyl inositol bi- ou triphosphate produit par la PI3K, mais que l'on retrouve mutée avec une perte de fonction dans de nombreux cancers [36].

b) L'insensibilité aux signaux d'inhibition de croissance

Tout comme pour les facteurs de croissance, il existe des inhibiteurs de croissance qui peuvent se présenter sous forme soluble, attachés à la matrice extracellulaire ou à la surface des cellules voisines. Et comme pour les facteurs de croissance, ces inhibiteurs de croissance sont perçus par la cellule grâce à des récepteurs transmembranaires qui transmettent le « message » à l'intérieur de la cellule. Ce message se traduit par l'entrée de la cellule soit dans la phase G_0 où elle ne se divise plus mais dont elle peut ressortir grâce aux signaux de croissance appropriés, soit dans une phase post-mitotique (généralement la différenciation de la cellule en un type cellulaire particulier) dont elle ne peut plus ressortir. Au niveau moléculaire, la plupart des inhibiteurs de croissance, comme le TGFβ (qui, contrairement à ce que laisse penser son nom, Tumor Growth Factor β, n'est pas un facteur de croissance comme le TGFα), induisent la déphosphorylation d'une protéine particulière, pRb (retinoblastoma protein) [37], qui est dès lors capable de séquestrer le facteur de transcription E2F. La prolifération cellulaire est ainsi bloquée car E2F est responsable de la transcription d'un grand nombre de gènes nécessaires à la progression dans le cycle cellulaire. Alors que la plupart des signaux externes à la cellule, devant arrêter sa progression dans le cycle cellulaire, activent pRb, les signaux internes activent p53. En effet, si l'ADN d'une cellule présente des dégâts ou si le niveau en nucléotides, en glucose ou en oxygène de cette cellule n'est pas optimal, p53 a la capacité d'activer différents effecteurs pour arrêter le cycle cellulaire jusqu'à un retour à un état homéostatique. Si les dégâts à l'ADN ne peuvent être réparés ou si les quantités en

nucléotides, en glucose ou en oxygène ne sont pas suffisantes au fonctionnement de la cellule, p53 peut même provoquer l'apoptose. Il faut aussi noter qu'il a été montré que des cellules souches embryonnaires de souris dépourvues soit de pRb [38], soit de p53 [39] ne montraient pas de prolifération anormale.

Au sein d'une cellule cancéreuse, la voie menant à la séquestration d'E2F peut être interrompue. Cela peut être la conséquence d'une diminution d'expression des récepteurs au TGFβ ou par la mutation d'une protéine de la cascade menant à la déphosphorylation de pRb [40]. C'est le cas, par exemple, des mutations des gènes de SMAD4 et CDK4 [41] ou de la délétion du gène de p15INK4B qui entraînent la phosphorylation de pRb, ce qui a pour conséquence la libération d'E2F et la progression dans le cycle cellulaire [42]. pRb peut lui-même être muté ou séquestré par une protéine d'origine virale (par exemple, la protéine E7 du virus du papillome humain [43]).

Enfin, en temps normal, les cellules (par exemple, les entérocytes et les érythroblastes) se différencient grâce au complexe des facteurs Mad et Max [44]. Max peut également s'associer à Myc pour éviter la différenciation, mais l'expression de Myc est bloquée par l'action du TGFβ. Cependant, si cette action est entravée par l'une des raisons expliquées avant, alors Myc peut supplanter Mad et s'associer à Max. Le complexe Myc-Max ainsi formé est un facteur de transcription permettant aux cellules cancéreuses d'éviter la différenciation et ainsi pouvoir continuer à se diviser [45].

La prolifération cellulaire est aussi régulée par le contact des cellules entre elles. C'est ce que l'on observe dans une boîte de culture avec des cellules non transformées : quand les cellules se sont divisées jusqu'à former une couche, les divisions s'arrêtent, et une seconde couche n'est jamais formée. On appelle ce phénomène l'inhibition de contact. Cependant, il ne s'observe pas dans tous les types cellulaires. Au niveau moléculaire, c'est le produit du gène NF2, qui séquestre des récepteurs de facteurs de croissance quand le nombre de contacts entre cellules devient important, qui est responsable de cette inhibition [46]. Inversement, pour les cellules cancéreuses en boîte de culture, on observe que les cellules sont capables de former plusieurs couches car l'inhibition de contact est perdue [47].

c) L'échappement à l'apoptose

Des conditions comme les dommages irréparables à l'ADN, l'insuffisance de facteurs de survie ou l'hypoxie (le manque d'oxygène) conduisent normalement la cellule à déclencher sa propre mort par apoptose [48]. Celle-ci se déroule en plusieurs étapes précises. D'abord, le

cytosquelette est démantelé, les chromosomes et le noyau sont fragmentés, et enfin, le contenu de la cellule est empaqueté dans des corps apoptotiques qui sont phagocytés notamment par des macrophages (Figure 3). Deux types de protéines permettent l'apoptose : les senseurs et les effecteurs. Les senseurs surveillent l'environnement extra- et intracellulaire. Ce sont, par exemple, des récepteurs transmembranaires comme IGF-1R, IL-3R ou TNF-R1 qui captent les facteurs de survie ou de mort IGF-1/2 (survie), IL-3 (survie) et TNFα (mort), respectivement. Les effecteurs, comme les caspases (signalons qu'il existe aussi des caspases qui sont des senseurs), sont des protéines qui détruisent la cellule selon les étapes citées précédemment si des signaux de mort ont été perçus par les senseurs. Bien sûr, à cela il faut ajouter toute une série d'acteurs qui permettent la transduction du signal. Citons, par exemple, les protéines adaptatrices qui se lient, notamment, aux récepteurs et aux caspases afin d'activer ces dernières.

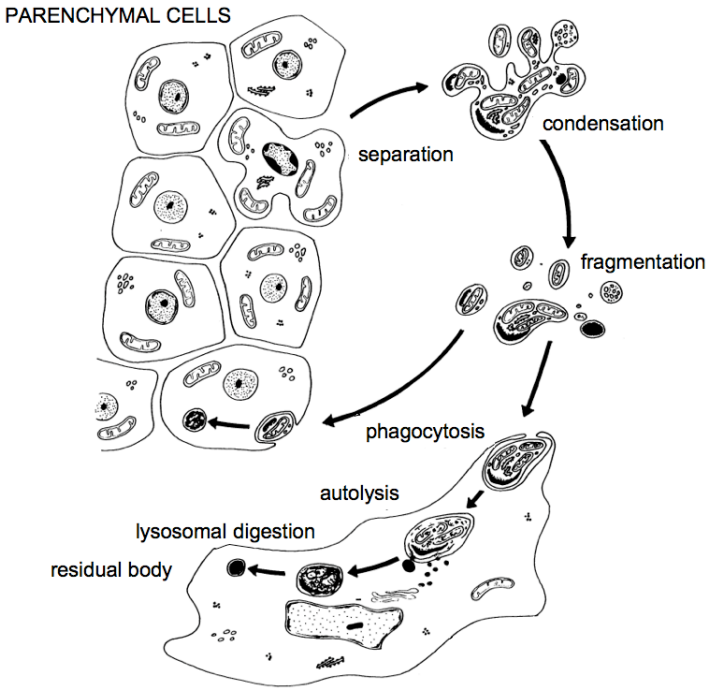


Figure 3. *Représentation schématique de l'apoptose (source : figure modifiée de [49]).*

Lors du déclenchement de l'apoptose, les signaux de mort conduisent à la libération de cytochrome c par la mitochondrie grâce aux protéines de la famille Bcl-2, dont les membres peuvent jouer un rôle pro-apoptotique (Bax et Bak) ou anti-apoptotique (Bcl-x_L, Bcl-w, Mcl-1 et A1). Au niveau moléculaire, les facteurs d'inhibition de l'apoptose de la famille Bcl-2 se lient aux facteurs pro-apoptotiques de la même famille, les empêchant ainsi d'induire l'apoptose (Figure 4). C'est donc l'équilibre entre les membres pro- et anti-apoptotiques des protéines de la famille Bcl-2 qui détermine si la cellule déclenche ou non l'apoptose [50]. Si

la proportion de facteurs pro-apoptotiques est plus grande, alors l'apoptose est déclenchée, et inversement. De plus, le suppresseur de tumeur p53, qui est capable de détecter les dommages à l'ADN, peut également conduire à l'activation des protéines de la famille Bcl-2 [51]. Une fois le cytochrome c libéré, celui-ci peut activer une protéase appelée caspase 9 qui, elle-même est capable d'activer toute une série d'autres caspases afin de réaliser physiquement l'apoptose [52]. Il est à noter que les récepteurs de FAS et du TNFα sont capables d'activer la caspase 8 dont la fonction est comparable à celle de la caspase 9. Enfin, signalons qu'il s'agit ici de l'apoptose canonique. Il existe d'autres mécanismes qui sont indépendants du cytochrome c ou des caspases, mais ceux-ci ne seront pas décrits dans ce travail.

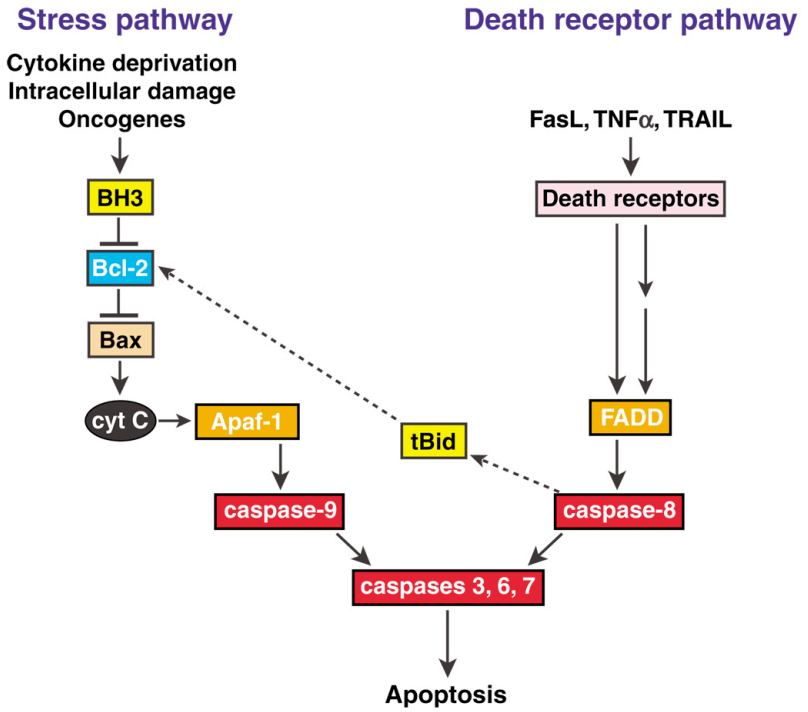


Figure 4. *La voie de stress est activée par une protéine de la famille BH3-only qui inactive Bcl-2 qui ne peut donc plus empêcher Bax et Bak d'induire la perméabilisation de la membrane externe de la mitochondrie menant ainsi à la libération de cytochrome c. Le cytochrome c permet alors à Apaf-1 d'activer la caspase 9 qui, elle-même, active d'autres caspases menant à l'apoptose. La voie des récepteurs de mort est, elle, activée quand des ligands, comme FAS et le TNFα, se lient à leur récepteur membranaire. Cela active la caspase 8, via la protéine adaptatrice FADD, engendrant ainsi l'apoptose. Il arrive que la voie des récepteurs de mort active la voie de stress via la protéine tBid (source : [50]).*

Plus de 50% des cancers humains présentent des mutations du gène de p53 [53]. Ceci résulte en la perte du plus puissant système de déclenchement de l'apoptose dont dispose la

cellule (Figure 5). De plus, pour éviter l'apoptose, les cellules cancéreuses présentent souvent une forte activité de la voie PI3 kinase–AKT/PKB, qui a pour rôle de transmettre des signaux anti-apoptotiques [54]. En outre, les cellules cancéreuses présentent souvent une expression élevée de facteurs anti-apoptotiques tels que Bcl-x_L ou de facteurs de survie comme Igf1/2, ainsi qu'une faible expression de facteurs pro-apoptotiques tels que Bax [55].

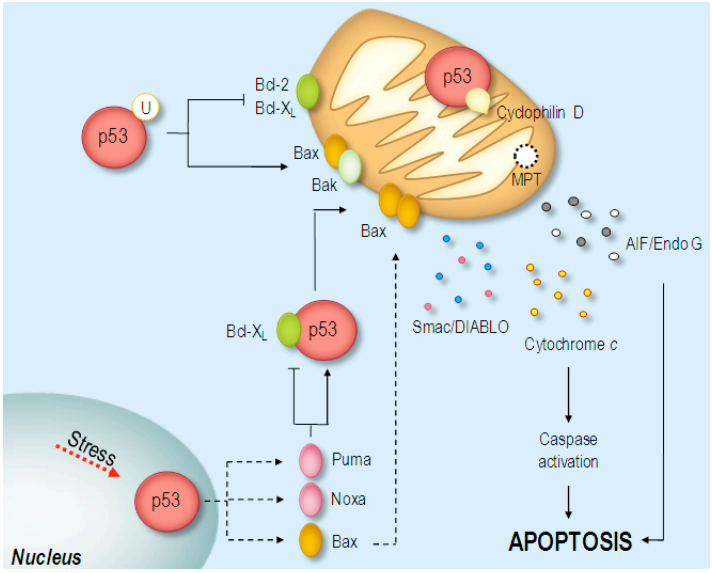


Figure 5. *Dans le noyau, p53 induit l'expression de Bax, de Noxa et de Puma. Ce dernier rend le p53 cytosolique actif en le libérant de Bcl-X_L. Dès lors, le p53 cytosolique induit l'oligomérisation de Bax et sa translocation au niveau de la mitochondrie. Dans le cytosol, p53 peut aussi porter un groupe ubiquitine lui permettant d'induire l'oligomérisation de Bax et de Bak et d'empêcher l'effet anti-apoptotique Bcl-2 et de Bcl-X_L au niveau de la mitochondrie. Dans celle-ci, p53 forme un complexe avec la cyclophiline D induisant la disruption des membranes mitochondriales. Ceci a pour conséquence la libération de facteurs pro-apoptotiques, dont le cytochrome c (source : [56], MPT = mitochondrial permeability transition, U = ubiquitine).*

d) Un potentiel de réplication illimité

Des cellules normales ne se répliquent pas indéfiniment. En effet, après un certain nombre de divisions, elles entrent dans un état de sénescence [57]. En invalidant pRb et p53, ces cellules peuvent encore se diviser plusieurs fois, jusqu'à arriver à un stade que l'on appelle « crise ». Cet état est caractérisé par un grand nombre de morts cellulaires dues à la fusion des extrémités chromosomiques [58]. Les chromosomes sont composés, à leurs extrémités, de télomères [59]. Les télomères sont constitués de milliers de répétitions de six nucléotides. À chaque réplication de l'ADN, qui se produit à chaque division de la cellule, les

télomères sont raccourcis de quelques dizaines de nucléotides car la protéine responsable de la réplication de l'ADN (l'ADN polymérase) est incapable de synthétiser entièrement l'un des brins de l'ADN car les chromosomes eucaryotes sont linéaires. Ceci résulte en un raccourcissement des chromosomes à chaque réplication. Quand ce raccourcissement a totalement éliminé les télomères et que ce sont les parties codantes de l'ADN qui sont touchées, c'est-à-dire celles qui permettent le fonctionnement de la cellule, les extrémités des chromosomes fusionnent, provoquant un chaos cellulaire aboutissant finalement à la mort de la cellule [60]. Cependant, parmi la population de cellules en crise, une sur dix millions peut acquérir la faculté de se diviser indéfiniment. Ce phénomène s'appelle l'immortalisation [61]. Il s'avère que de nombreuses cellules cancéreuses sont immortalisées. En effet, celles-ci surexpriment les télomérases. Les télomérases sont des enzymes capables d'ajouter des hexanucléotides aux télomères [62]. Cependant, hormis dans les cellules germinales, elles sont peu actives. Surexprimées, les télomérases permettent l'élongation des télomères et ainsi la possibilité, pour la cellule, de se répliquer indéfiniment puisque les parties codantes de son ADN ne seront jamais touchées par le raccourcissement des extrémités chromosomiques.

e) La néoangiogenèse

Toute cellule a besoin d'être proche (au maximum à 100 µm) d'un vaisseau sanguin afin d'obtenir des quantités suffisantes en oxygène et en nutriments et afin d'évacuer les déchets du métabolisme et le dioxyde de carbone. Durant la formation des organes, la mise en place du réseau vasculaire se fait par le phénomène d'angiogenèse. Celui-ci peut encore être déclenché plus tard, mais nécessite la production de facteurs angiogéniques au préalable. En effet, pour enclencher leur prolifération, les cellules qui composent les vaisseaux sanguins, dont les cellules endothéliales, doivent percevoir, via des récepteurs appropriés, des facteurs tels que le VEGF (Vascular Endothelial Growth Factor) ou le FGF (Fibroblast Growth Factor) [63]. À l'inverse, il existe aussi des facteurs inhibiteurs de l'angiogenèse comme la thrombospondine-1 [64].

Au sein des tumeurs, les cellules cancéreuses opèrent un changement dans la balance des activateurs et des inhibiteurs de l'angiogenèse [65]. En effet, il n'est pas rare d'observer dans les cellules cancéreuses une augmentation de l'expression de VEGF et/ou de FGF et une diminution de l'expression de la thrombospondine-1. L'augmentation de l'expression de VEGF peut notamment être due à l'activation de Ras, qui a déjà été décrit plus haut comme étant impliqué dans le déclenchement de la mitose des cellules cancéreuses, ou par l'hypoxie [66]. En effet, à mesure que la tumeur grandit, l'oxygène manque de plus en plus en son

centre. Cela a pour effet d'activer le facteur de transcription HIF-1 (Hypoxia-Inducible Factor-1), dont l'une des cibles est le gène de VEGF. La sous-expression de la thrombospondine-1 peut être due à la perte de fonction de p53 [67]. En effet, p53, en plus de déclencher l'apoptose, est responsable de la régulation de l'expression de gènes comme celui de la thrombospondine-1 [68].

Ces différents évènements ont pour conséquence la mise en place d'un nouveau système vasculaire au sein des tumeurs (Figure 6). Cependant, celui-ci possède une structure aberrante (Figure 7). En effet, les ramifications apparaissent de manière précoce et en trop grand nombre. Les vaisseaux sont soit trop étroits, soit trop larges. De plus, leur étanchéité n'est pas parfaite puisqu'il n'est pas rare d'observer des microhémorragies [69]. Les flux sanguins sont donc irréguliers au sein des tumeurs.

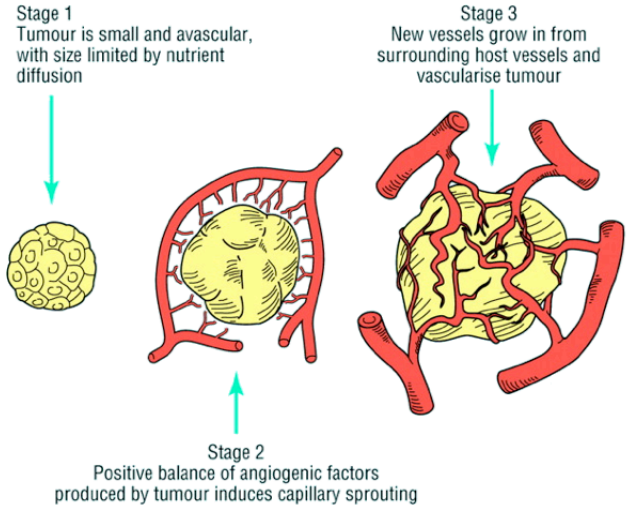


Figure 6. *Les étapes de la néoangiogenèse (source : [70]).*

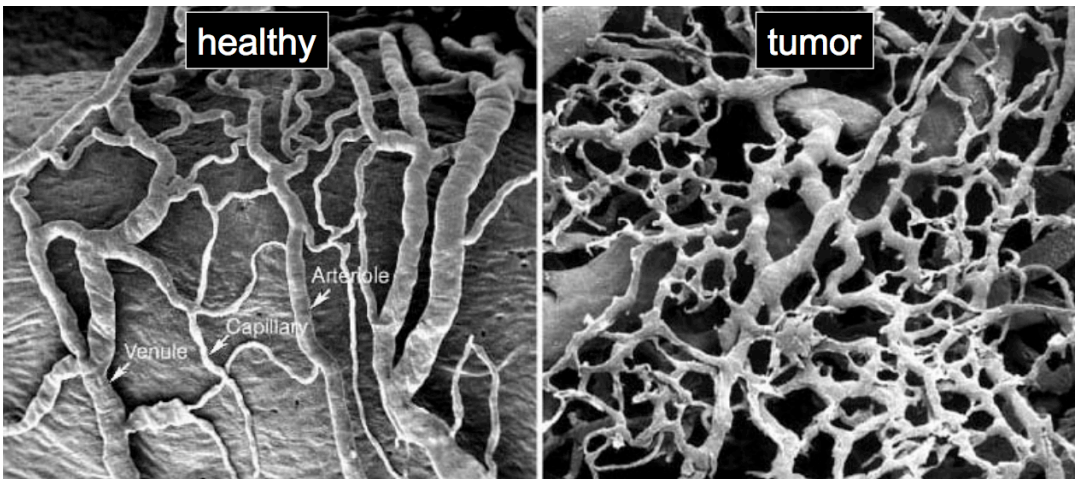


Figure 7. *Vaisseaux sanguins normaux (à gauche) et tumoraux (à droite) visualisés en microscopie électronique à balayage après moulage (source : [71]).*

f) Les métastases

Tôt ou tard dans le développement de la plupart des cancers humains, une ou plusieurs cellules se détachent de la tumeur primaire pour envahir les tissus adjacents et circuler vers des sites distaux de l'organisme où, au début en tout cas, l'espace et les nutriments sont en abondance [72]. Les métastases sont la cause de 90% de décès par le cancer. Étant le sujet de ce travail, les métastases et leurs caractéristiques seront décrites plus en détails dans le chapitre suivant.

g) La reprogrammation du métabolisme énergétique

Comme expliqué plus haut, les cellules cancéreuses se trouvent souvent dans un milieu pauvre en oxygène. Ceci est dû à l'éloignement des vaisseaux sanguins et au mauvais fonctionnement du nouveau réseau circulatoire formé au sein des tumeurs. C'est pourquoi, la glycolyse est favorisée aux dépens de la phosphorylation oxydative, qui nécessite de l'oxygène, au sein des cellules cancéreuses. Cependant, bien que la glycolyse ait un rendement, dans la production d'énergie, beaucoup moins élevé que la phosphorylation oxydative, les cellules cancéreuses continuent de favoriser la glycolyse même en présence d'oxygène [73]. Ceci est rendu possible par une expression accrue de GLUT1, qui se trouve à la surface des cellules et dont l'une des fonctions est d'importer le glucose dans la cellule [74]. De plus, l'activation du facteur de transcription HIF-1, par la protéine Ras et l'hypoxie, augmente encore la vitesse de la glycolyse, puisque de nombreux autres transporteurs de glucose et des enzymes de la glycolyse sont encodées par des gènes cibles de HIF-1 [75].

Selon une ancienne hypothèse [76], qui a été reconsidérée récemment, les cellules cancéreuses opéreraient cette reprogrammation du métabolisme énergétique car la glycolyse, bien que moins efficace, produit aussi des intermédiaires qui peuvent être utilisés dans la synthèse des nucléosides et des acides aminés. Ceux-ci peuvent alors servir à la synthèse des macromolécules et des organites nécessaires à la formation des nouvelles cellules [77].

Il est aussi intéressant de noter qu'il existe au sein de certaines tumeurs deux sous-populations de cellules cancéreuses : celles qui sont peu oxygénées et celles qui sont bien oxygénées. Les cellules cancéreuses peu oxygénées, qui opèrent la glycolyse, rejettent du lactate qui est ensuite utilisé par les cellules cancéreuses mieux oxygénées comme source principale d'énergie. En effet, le lactate peut être employé dans le cycle de l'acide citrique pour produire de l'énergie [78].

h) L'échappement au système immunitaire

On a longtemps pensé que le système immunitaire surveillait l'entièreté de l'organisme et éliminait la plupart des tumeurs naissantes. Les tumeurs qui, malgré cela, arrivaient à se développer avaient réussi à éviter la surveillance du système immunitaire ou à limiter son action. On observe, d'ailleurs, chez des souris déficientes pour certaines cellules du système immunitaire (lymphocytes T cytotoxiques CD8+, cellules T CD4+ « Th1 helper » et « Natural Killer ») qu'elles développent plus et/ou plus rapidement des tumeurs [79]. De la même manière, il a été montré que des patients atteints de cancers du colon ou des ovaires présentaient un meilleur pronostic si les tumeurs étaient infiltrées par un grand nombre de lymphocytes T cytotoxiques CD8+ ou de « Natural Killers » [12, 80].

Cependant, on s'aperçoit que les cellules cancéreuses sont, tout de même, capables de résister au système immunitaire. Par exemple, les cellules cancéreuses peuvent empêcher l'action des lymphocytes T cytotoxiques CD8+ et des « Natural Killers » en sécrétant du TGF-β ou d'autres facteurs immunosuppresseurs [81], ou en recrutant d'autres cellules inflammatoires ayant un effet immunosuppresseur [82].

Le cancer est donc une maladie dont le développement est long et complexe, mais qui peut être expliqué par huit capacités qu'acquièrent les cellules suite à l'instabilité du génome, à l'inflammation et à l'influence de leur microenvironnement. Le cancer peut toucher un grand nombre de types cellulaires, et donc un grand nombre de tissus et d'organes, et peut être provoqué par de nombreux facteurs tant exogènes qu'endogènes.

2.2. Les métastases

2.2.1. Généralités

Les métastases sont responsables de la plupart des décès causés par le cancer. Elles représentent l'étape ultime dans le développement cancéreux. Il s'agit d'un processus complexe se déroulant en plusieurs étapes. Ces différentes étapes sont causées par des mécanismes moléculaires et des changements dans le profil d'expression génique des cellules cancéreuses [83] qui sont au centre de ce travail. Il est, toutefois, important de noter que de nombreuses données concernant l'étude des métastases sont, encore aujourd'hui, mal comprises et parfois contradictoires [84]. Cependant, ce chapitre va tenter de décrire brièvement les différentes étapes de la cascade métastatique qui sont communément admises.

Pour métastasier, une ou plusieurs cellules cancéreuses doivent se détacher de la tumeur primaire. Celles-ci doivent, ensuite, envahir les tissus environnants jusqu'à pénétrer un vaisseau sanguin ou lymphatique. Cette étape permet aux cellules cancéreuses de voyager dans tout l'organisme. L'étape suivante est l'arrêt et la sortie du système circulatoire qui est suivie de la colonisation de l'organe ainsi atteint [85].

Les cellules cancéreuses qui métastasient sont soumises, à chacune de ces étapes, à une élimination massive. En effet, que ce soit les cellules du système immunitaire ou la force du flux sanguin dans les vaisseaux, les difficultés pour une cellule à métastasier ne manquent pas. C'est pour cela que, malgré la grande quantité de cellules qui se détachent de la tumeur primaire, seules quelques-unes parviennent à métastasier avec succès. De plus, nombreuses sont les cellules métastatiques qui, une fois arrivées au site secondaire, entrent en dormance pendant une période de parfois plusieurs années avant d'être, à nouveau, actives.

2.2.2. La cascade métastatique

a) La transition épithéliale-mésenchymale

À la surface des cellules épithéliales, se trouvent des jonctions intercellulaires. Celles-ci permettent à l'épithélium de maintenir son intégrité et aux cellules de garder leur polarité. Parmi ces jonctions intercellulaires, l'on retrouve celles médiées par les E-cadhérines (Epithelial cadherins). Les E-cadhérines sont des molécules transmembranaires dont le domaine extracellulaire interagit avec des molécules de E-cadhérine d'autres cellules voisines (il s'agit d'interactions homophiliques) et dont le domaine cytoplasmique se lie à des protéines de la famille des caténines, qui permettent la liaison à l'actine du cytosquelette. De plus, en liant la β-caténine, l'E-cadhérine limite sa quantité dans le cytosol favorisant ainsi la survie cellulaire puisque la β-caténine est un facteur pouvant induire l'apoptose [86].

La première étape qu'une cellule cancéreuse doit réaliser pour migrer est de se détacher de la tumeur primaire. Cela se produit en diminuant ou en éliminant l'expression de l'E-cadhérine (Figure 8). On observe, d'ailleurs, fréquemment la perte de l'E-cadhérine au niveau des cellules du front invasif de certains cancers du sein [87], du rein [88] ou de l'ovaire [89]. De même, *in vitro*, on observe que la disruption de *CDH1* (le gène codant pour l'E-cadhérine) dans des cellules non invasives les rend invasives. Et, au contraire, la surexpression de l'E-cadhérine dans des cellules invasives leur fait perdre cette capacité [90].

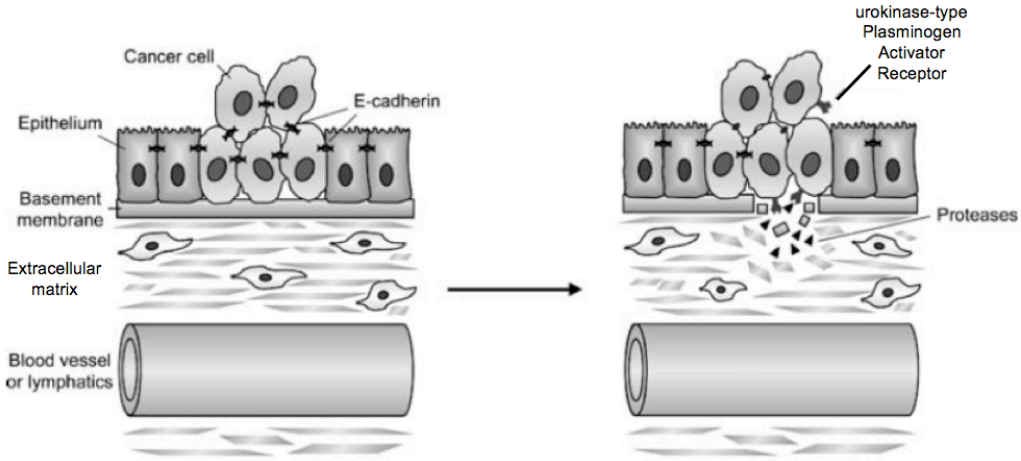


Figure 8. *La première étape pour que des cellules cancéreuses métastasient est de rompre les contacts avec les cellules voisines. Cela se fait notamment par répression de l'expression de l'E-cadhérine (source : figure modifiée de [85]).*

Enfin, la perte de l'expression de l'E-cadhérine est souvent accompagnée d'une augmentation de l'expression de la N-cadhérine (Neural cadherin). Ceci est dû au fait que la perte de l'E-cadhérine entraîne une augmentation de l'activité de NFκB (Nuclear Factor kappa B) qui est un facteur de transcription dont l'une des cibles est le gène codant pour la N-cadhérine. En effet, la perte d'expression de l'E-cadhérine entraîne une augmentation de la quantité de la β-caténine dans le cytosol, celle-ci active alors p38 qui active à son tour NFκB [91]. Bien que la N-cadhérine soit capable, comme l'E-cadhérine, d'interactions homophiliques, elle est aussi capable d'interactions hétérophiliques, c'est-à-dire avec d'autres molécules, comme la placoglobine [92], exprimées sur d'autres types cellulaires, par exemple des cellules endothéliales et des fibroblastes [93]. Cette capacité permet aux cellules de migrer à travers les tissus qui contiennent des cellules endothéliales et des fibroblastes.

Il faut toutefois noter que le modèle de la transition épithéliale-mésenchymale reste hypothétique. Bien que de plus en plus de preuves s'accumulent en sa faveur, certains auteurs [94-97] ont montré que des cellules qui opèrent la transition épithéliale-mésenchymale ne prolifèrent pas, remettant dès lors en cause que ce mécanisme est mis en place dans des processus comme la formation du mésoderme ou du tube neural ou dans des pathologies comme le cancer.

b) L'envahissement de la membrane basale et de la matrice extracellulaire

Une fois détachées de la tumeur primaire, les cellules cancéreuses qui métastasient doivent encore passer la barrière que constituent la membrane basale et la matrice

extracellulaire. La seconde étape de la cascade métastatique est donc la dégradation de cette barrière qui sépare la tumeur des tissus voisins. Cette dégradation nécessite l'expression de toute une série d'enzymes protéolytiques puisque la membrane basale et la matrice extracellulaire sont, principalement, composées de protéines.

Les cathepsines font partie de ces protéases. On en connaît, à l'heure actuelle, une douzaine qui présentent différents mécanismes catalytiques, différentes structures et qui ont différentes cibles. Elles sont actives dans les lysosomes où le pH est optimal pour leur fonctionnement [98]. Cependant, de plus en plus d'études montrent que les cathepsines B, L et D, notamment, joueraient un rôle important dans le développement des tumeurs malignes. Celles-ci seraient sécrétées par les cellules cancéreuses et dégraderaient la matrice extracellulaire [99]. De plus, outre sa capacité protéolytique, la cathepsine D permettrait la division cellulaire et diminuerait la réponse immunitaire dirigée contre les cellules cancéreuses en inhibant la fonction des cellules dendritiques [100].

Les MMPs (matrix metalloproteinases) forment une famille d'enzymes protéolytiques exerçant un grand nombre de fonctions. Elles sont notamment impliquées dans des processus physiologiques normaux comme le développement, la morphogenèse ou la réparation de tissus [101]. Tous ces processus de remodelage demandent une dégradation précisément régulée de la matrice extracellulaire (MEC). Cependant, le remodelage intervient également dans diverses maladies dont le cancer, les mécanismes de dégradation de la MEC sont alors souvent dérégulés [102]. Les MMPs peuvent non seulement contribuer à l'invasion du cancer par la dégradation de la MEC mais aussi en libérant des facteurs de croissance qui étaient séquestrés [103].

D'autre part, l'envahissement de la MEC est aussi possible grâce à l'uPA (urokinase-type Plasminogen Activator). uPA se lie à son récepteur, uPAR (urokinase-type Plasminogen Activator Receptor), à la surface des cellules. L'uPA, ainsi activé, transforme le plasminogène en plasmine qui est capable de dégrader plusieurs protéines de la matrice extracellulaire (par exemple, la fibronectine [104]) ainsi que d'activer des facteurs de croissance (par exemple, le VEGF [105]) et des métalloprotéinases (par exemple, MMP3 [106]) (Figure 9).

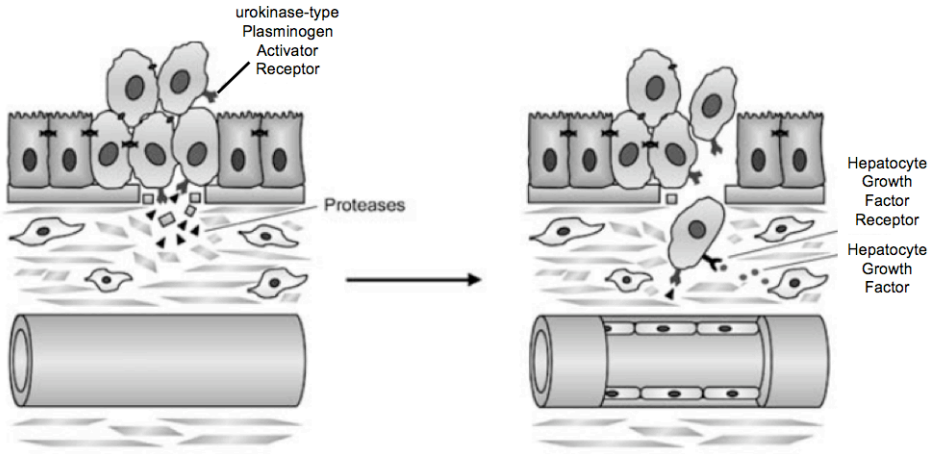


Figure 9. *Rupture de la membrane basale et de la matrice extracellulaire (source : figure modifiée de [85]).*

La migration cellulaire à proprement parlé est aussi régulée par uPAR. En effet, uPAR a la capacité de se lier à des protéines de surface d'autres types cellulaires telles que des intégrines, des récepteurs de facteurs de croissance ou des récepteurs de facteurs de migration [107]. Enfin, en se liant à la vitronectine, uPAR rend la cellule adhérente à la matrice extracellulaire [108] ; alors qu'en liant les intégrines β1 ou β2, il fait perdre cette adhérence à la cellule [109]. C'est donc le cycle de gain et de perte d'adhérence de la cellule aux autres cellules ou à la matrice extracellulaire qui lui permet d'avancer physiquement dans son environnement.

c) La migration cellulaire

Les effets des protéases extracellulaires n'expliquent pas à eux seuls la migration des cellules cancéreuses qui métastasient. En effet, afin de pénétrer les tissus, les cellules tumorales utilisent aussi les mêmes mécanismes de migration que les cellules normales lors de la morphogenèse embryonnaire ou du trafic des cellules immunitaires. La migration des cellules se fait en plusieurs étapes. D'abord, les cellules se polarisent et s'allongent. Ensuite, un pseudopode se forme et s'attache à la matrice extracellulaire. Enfin, le corps cellulaire se contracte [83].

L'élongation des pseudopodes se fait par polymérisation d'actine jusqu'à ce qu'il y ait contact avec la matrice extracellulaire. Les interactions entre les pseudopodes et la matrice extracellulaire se font via des molécules d'adhérence, dont des récepteurs de type intégrine (par exemple, les intégrines 1 et 3 [110]), pour former des complexes focaux qui se développent et se stabilisent en contacts focaux. La présence de récepteurs d'adhérence entraîne le recrutement de protéases de surface (par exemple, les métalloprotéinases MMP1, 2

et 9 [111]) qui dégradent la matrice extracellulaire ce qui crée un gradient permettant à la cellule de s'orienter. Durant le développement des contacts focaux, des filaments d'actine s'assemblent au sein de la cellule. Ces filaments se contractent grâce à la myosine II, ce qui provoque le rétrécissement de l'arrière de la cellule permettant sa progression [112].

d) L'intravasion et l'extravasion

Une fois arrivées à proximité d'un vaisseau sanguin ou lymphatique, les cellules cancéreuses migratoires doivent encore pénétrer à l'intérieur de ce vaisseau afin de poursuivre la cascade métastatique. Elles y parviennent en surexprimant le gène du VEGF. Le VEGF augmente la perméabilité du réseau vasculaire [113] en rompant la barrière que forment les cellules endothéliales des vaisseaux, ceci en brisant les jonctions qui les lient, en réarrangeant l'actine des cellules et en créant des vides entre elles. Ce phénomène s'appelle l'intravasion. De plus, des études ont montré que la surexpression du VEGF était positivement corrélée à une pression des fluides interstitiels élevée au sein des tumeurs, ce qui facilite encore davantage la pénétration des cellules cancéreuses migratoires, qui sont littéralement poussées dans les vaisseaux [114].

Lors de l'extravasion (la sortie des cellules des vaisseaux), c'est également le VEGF qui permet aux cellules se trouvant dans la circulation de sortir des vaisseaux en augmentant leur perméabilité comme pour l'intravasion [115].

Une fois sorties de la circulation à un organe distant de la tumeur primaire, seules les cellules capables de stimuler l'angiogenèse en produisant du VEGF ou se trouvant à proximité d'un vaisseau sanguin pourront induire le développement d'une tumeur secondaire (Figure 10) [116]. Les autres resteront en dormance jusqu'à pouvoir éventuellement redevenir actives si d'autres cellules se trouvant dans l'environnement produisent du VEGF [117].

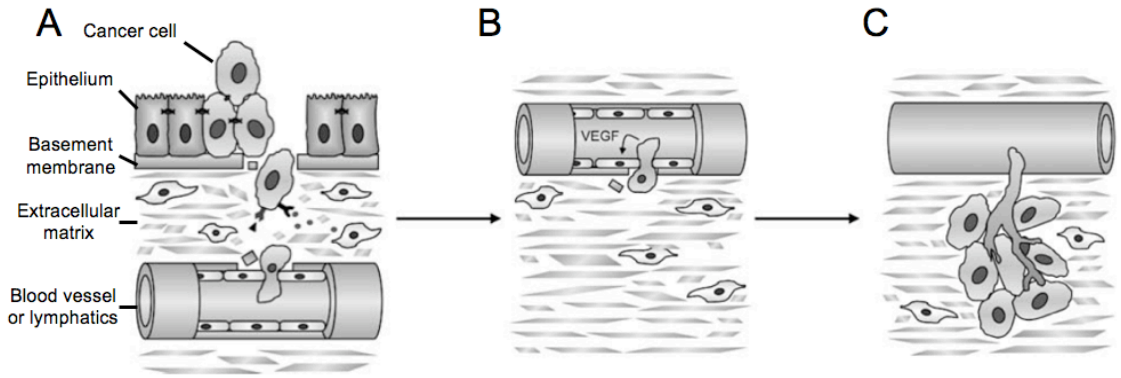


Figure 10. *(**A**) Intravasion et (**B**) extravasion des cellules cancéreuses et (**C**) développement d'une tumeur secondaire (source : figure modifiée de [85]).*

2.2.3. Métastases et hypoxie

Un élément clé du développement des tumeurs est la présence de zones hypoxiques en leur centre. Les régions hypoxiques au sein d'une tumeur sont le résultat de l'augmentation progressive de la distance entre les cellules et les vaisseaux sanguins lors du développement de la tumeur. De plus, bien que les cellules cancéreuses enclenchent un processus de néo-vascularisation par angiogenèse, le réseau sanguin nouvellement formé est anormal et ne permet qu'un apport irrégulier d'oxygène et de nutriments aux cellules (Figure 11). Les patients dont des zones tumorales se trouvent en hypoxie développent dans la majorité des cas des cancers agressifs. Plusieurs hypothèses ont été proposées pour expliquer cette observation :

- l'hypoxie déclenche une adaptation résultant de l'activation du facteur de transcription HIF-1 qui améliore la survie des cellules cancéreuses [118],
- cette adaptation déclenche également le processus d'angiogenèse,
- l'hypoxie conduit à une résistance des cellules cancéreuses aux radiothérapies et aux chimiothérapies [119],
- de plus en plus de données expérimentales suggèrent que l'hypoxie améliore et/ou sélectionne les cellules cancéreuses à haut potentiel métastatique [120-122].

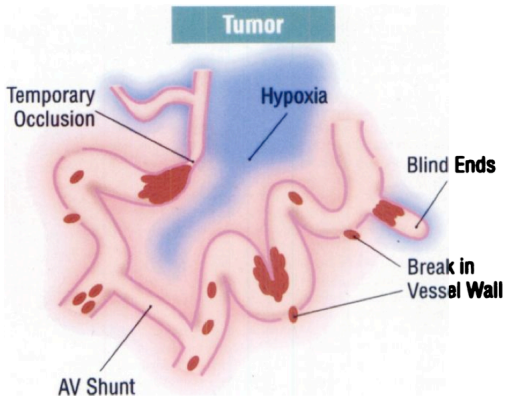


Figure 11. *Schéma représentant une zone d'hypoxie au sein d'une tumeur due à l'éloignement des cellules cancéreuses par rapport aux vaisseaux sanguins et à la structure anormale de ces vaisseaux (source : [123]).*

Cette section va revenir sur la cascade métastatique, précédemment décrite, en tentant de mettre en lumière le rôle que joue l'hypoxie dans son déroulement. En effet, de nombreux mécanismes moléculaires mis en place lors de la cascade métastatique sont, en fait, déclenchés, en partie ou exclusivement, par le manque d'oxygène.

a) La transition épithéliale-mésenchymale

La perte de l'expression de l'E-cadhérine est souvent observée au niveau du front invasif des tumeurs. La perte de cette jonction intercellulaire permet aux cellules cancéreuses de se détacher de la tumeur primaire. Il est rare que la perte de l'E-cadhérine soit le résultat d'une mutation dans son gène. Il semblerait que ce soit l'hypoxie qui en soit responsable. En effet, plusieurs études ont montré une diminution d'expression de l'E-cadhérine en condition hypoxique, ainsi qu'une expression mutuellement exclusive de l'E-cadhérine et de HIF-1α [88-90, 124-126]. HIF-1α est l'une des sous-unités du facteur de transcription HIF-1, qui permet une adaptation de la cellule au manque d'oxygène. En normoxie, HIF-1α est hydroxylé par une PHD (prolyl-4-hydroxylase) sur les prolines 402 ou 564. Ceci conduit à favoriser une interaction de HIF-1α hydroxylé avec le complexe VHL (von Hippel-Lindau) qui agit comme une E3 ubiquitine ligase, c'est-à-dire qu'il lui fixe des molécules d'ubiquitine, le transformant ainsi en une cible du protéasome qui se chargera de le dégrader. Par contre, en absence d'oxygène, PHD ne peut fonctionner. HIF-1α ne peut donc être dégradé. Il transloque alors dans le noyau où il s'associe avec la sous-unité β pour former le facteur de transcription fonctionnel HIF-1, qui permet la transcription de gènes cibles (Figure 12) [118].

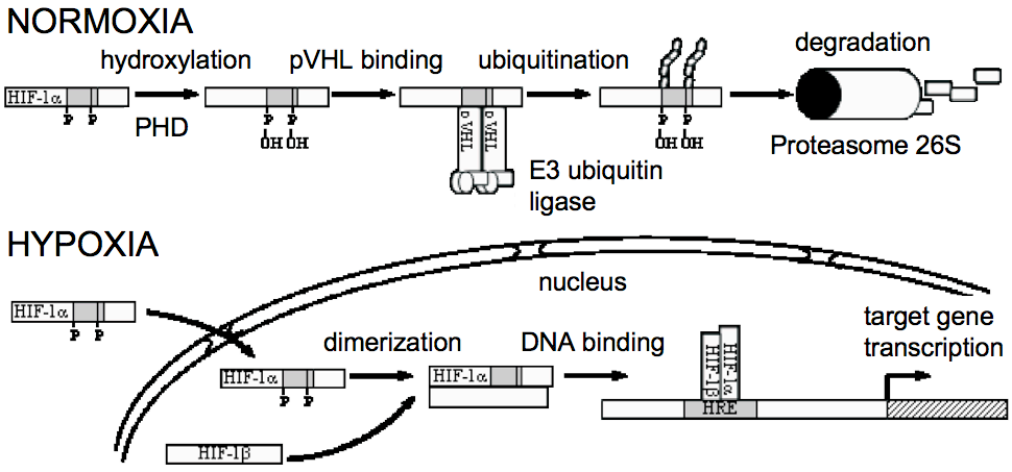


Figure 12. *En normoxie, HIF-1α est hydroxylé par une PHD (prolyl-4-hydroxylases) puis des groupements ubiquitines lui sont ajoutés par le complexe VHL (von Hippel-Lindau) qui entraînent sa dégradation par le protéasome. En hypoxie, HIF-1α n'est pas hydroxylé, il peut ainsi pénétrer dans le noyau où il s'associe avec HIF-1β pour former HIF-1 (source : [127]).*

Plusieurs répresseurs transcriptionnels sont capables de bloquer l'expression de l'E-cadhérine. Parmi ceux-ci, l'on retrouve SLUG, TCF3, ZFHX1A, ZFHX1B et SNAIL. Des études ont montré que l'expression de SLUG était augmentée en condition hypoxique [125]. De même, il semble que des cellules de cancer du rein, dont l'expression de HIF-1α est

bloquée, sont incapables de synthétiser TCF3, ZFHX1A et ZFHX1B. Enfin, l'expression de SNAIL est augmentée par le facteur AMF (Autocrine Motility Factor) dont l'expression est sous le contrôle de HIF-1 [128].

b) L'envahissement de la membrane basale et de la matrice extracellulaire

Afin de dégrader la membrane basale et la matrice extracellulaire, les cellules cancéreuses qui métastasient doivent surexprimer uPAR. Ici encore, l'hypoxie joue un grand rôle puisque uPAR est sous le contrôle transcriptionnel de HIF-1. En effet, des sites de liaison de HIF ont été localisés en amont du promoteur de uPAR [129]. Une surexpression de uPAR et une augmentation des capacités invasives ont, par ailleurs, été rapportées dans des cancers du sein, du colon, de la prostate et dans des mélanomes [130]. Enfin, une surexpression de HIF-1α ou une sous-expression du complexe VHL, *in vitro*, est positivement corrélée à une augmentation de la quantité de uPAR, alors qu'une invalidation de HIF-1α par siRNA entraîne une diminution du taux d'uPAR [130].

En plus d'être sous le contrôle de HIF-1, il semble que la régulation d'uPAR soit également dépendante d'autres voies de signalisation, telles que la voie des MAPK, qui, elle-même, est activée par l'hypoxie [131].

c) L'intravasion et l'extravasion

Lors de l'intravasion et de l'extravasion, les cellules en migration produisent du VEGF afin d'augmenter la perméabilité des vaisseaux sanguins. Une fois encore, c'est l'hypoxie qui est principalement responsable de la surexpression de ce facteur. En effet, le promoteur du gène codant pour le VEGF possède plusieurs HRE (Hypoxia Response Element) qui sont les sites de liaison pour HIF-1 [132].

Ensuite, rappelons que le VEGF est responsable de la néoangiogenèse et donc offre aux cellules cancéreuses de la tumeur primaire de nouvelles voies d'accès pour atteindre la circulation sanguine de l'organisme. L'hypoxie joue donc un rôle favorable aux métastases à ce niveau également.

2.3. Conclusion de la problématique biologique

Les cancers sont donc des maladies provoquées par des mutations de gènes particuliers : les oncogènes, les gènes suppresseurs de tumeur et les gènes de réparation de l'ADN. Plusieurs mutations sont souvent nécessaires pour entraîner la maladie, mais chacune

facilite l'apparition de la suivante. Ainsi, les mutations apparaissent selon une séquence aléatoire.

La transformation d'une cellule normale en une cellule cancéreuse résulte de l'acquisition par celle-ci de plusieurs capacités, elles-mêmes dues aux mutations qu'a subi l'ADN. Ces capacités permettent aux cellules cancéreuses :

- de se multiplier sans en avoir reçu le signal,
- d'être insensibles aux signaux les empêchant de se diviser,
- d'empêcher leur mort programmée,
- de se multiplier à l'infini,
- de stimuler la néovascularisation afin d'obtenir l'oxygène et les nutriments nécessaires à leur prolifération et à leur survie,
- de passer à un métabolisme anaérobique,
- d'échapper au système immunitaire,
- de métastasier pour envahir des sites distaux.

Les métastases sont le résultat de plusieurs étapes successives : détachement de la tumeur primaire, dégradation de la matrice extracellulaire, migration, intravasion et extravasion. De nombreux mécanismes moléculaires de cette cascade métastatique sont déclenchés par le manque d'oxygène que l'on observe au sein des tumeurs. En effet, l'hypoxie entraîne des changements, notamment au niveau transcriptionnel, qui permettent aux cellules cancéreuses de s'adapter au manque d'oxygène et de fuir ce milieu hostile.

Les connaissances de ces changements sont encore incomplètes et toute nouvelle découverte constituerait une précieuse avancée dans la compréhension du processus métastatique et permettrait de cibler de nouveaux facteurs afin de lutter contre ces maladies qui tuent encore trop de patients à l'heure actuelle.

3. La problématique méthodologique

Apparues il y a plus de dix ans dans le monde de la recherche en biologie moléculaire, les puces à ADN sont devenues l'un des outils majeurs utilisés dans beaucoup de laboratoires. Cette technique permet de mesurer, en une seule expérience, l'abondance des différents ARNm se trouvant, à un instant donné, dans un échantillon biologique [133]. Cependant,

malgré les perspectives que laissait entrevoir cette nouvelle technique à haut débit, les puces à ADN ont, depuis leurs débuts, dû faire face à de nombreuses critiques venant de la communauté scientifique. En effet, tant au niveau de leur conception que de l'analyse des résultats, les puces à ADN posent de nombreux problèmes.

Toutefois, la communauté scientifique, en particulier par des approches bioinformatiques et biostatistiques, a continué de développer les techniques permettant une analyse efficace des résultats provenant de puces à ADN. Malgré l'avènement de techniques à haut débit censées remplacer les puces à ADN, il est aujourd'hui possible d'émettre de nouvelles hypothèses dans un grand nombre de problématiques biologiques grâce aux résultats obtenus à l'aide des puces à ADN.

Ce chapitre va tenter d'expliquer ce que sont les puces à ADN, comment les résultats que l'on en obtient sont analysés et comment ils permettent de générer de nouvelles pistes pour des validations par des techniques *in vitro* et *in vivo*.

3.1. Les puces à ADN

3.1.1. Contexte biologique

Comme cela a été vu dans la partie précédente, traitant de la problématique biologique, tout organisme est constitué de cellules qui contiennent de l'ADN. Certaines séquences d'ADN forment des gènes qui permettent à la cellule, après transcription et traduction, de réaliser ses différentes fonctions. Bien sûr, il s'agit là d'une simplification puisque certains gènes ne codant pas pour des protéines, il n'y a dès lors pas de traduction.

Toute fonction n'étant pas requise en permanence, il existe, au sein des cellules, divers systèmes de régulation permettant de contrôler l'expression des gènes. Que ce soit au niveau transcriptionnel, post-transcriptionnel, traductionnel ou post-traductionnel, il existe une grande variété de mécanismes permettant d'activer ou de bloquer les acteurs moléculaires de la cellule.

Ici, nous ne nous intéresserons qu'au niveau transcriptionnel de la régulation. Comme décrit auparavant, l'ADN doit être transcrit en ARNm qui sera, lui-même, traduit en protéine. Il existe, chez les eucaryotes, une étape supplémentaire précédant la traduction : leurs gènes sont composés d'introns (séquences non codantes) et d'exons (parties transcrites des gènes qui codent pour des protéines). Cette mosaïque permet, à partir d'un seul gène, de générer plusieurs ARNm, et *in fine* plusieurs protéines, par les différentes combinaisons possibles

d'exons. Les combinaisons nécessitent l'épissage des introns. Mais avant cette étape, une coiffe et une queue polyA sont ajoutées aux extrémités de l'ARN pré-messager afin de lui permettre de sortir du noyau une fois l'épissage effectué [134]. Lors de cette étape, un complexe ribonucléoprotéique, le spliceosome, excise les introns et ligature les exons qui forment alors l'ARN mature [135] (Figure 13).

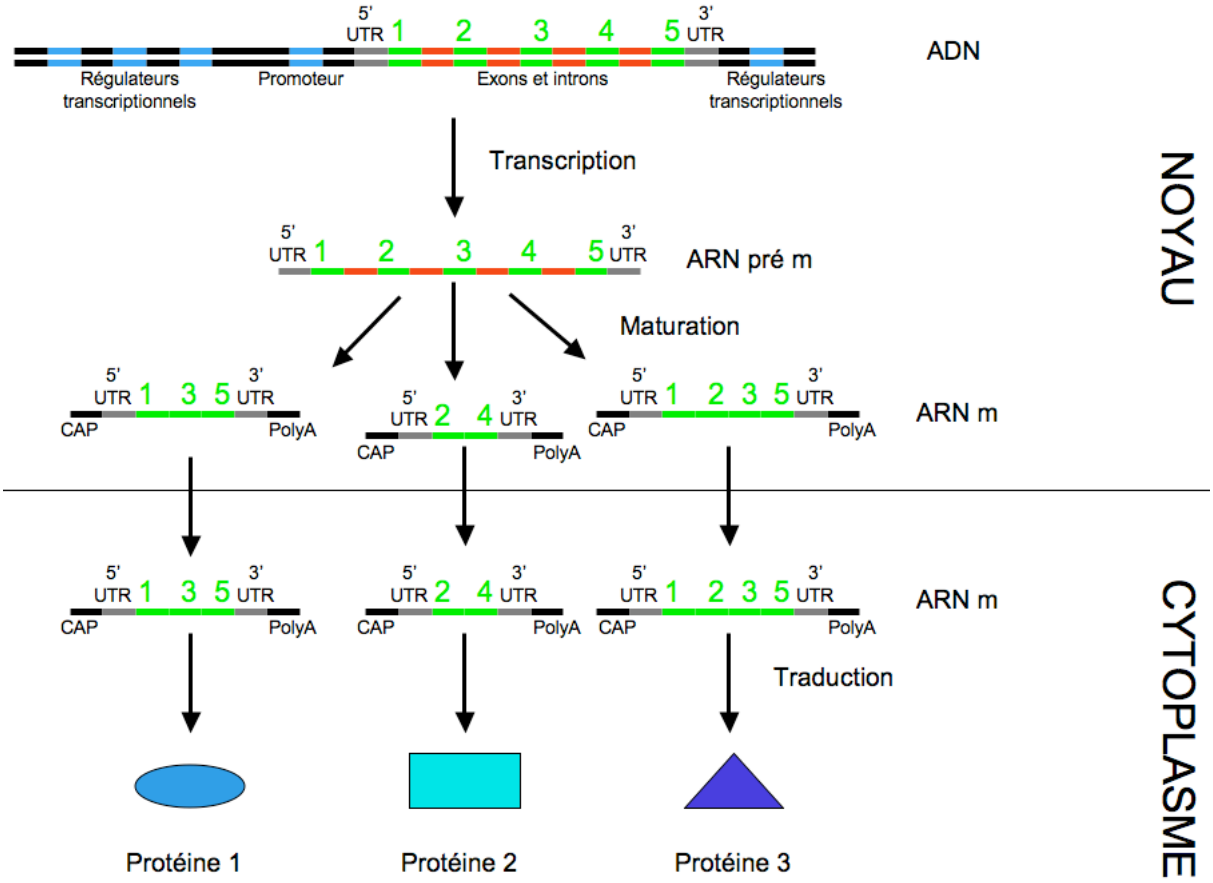


Figure 13. *Chez les eucaryotes, après transcription, l'ARN pré-messager doit encore être maturé avant de subir la translocation hors du noyau pour pouvoir être traduit. Cette maturation implique l'excision des introns (en rouge) pour ne laisser que des combinaisons d'exons (en vert). Ceci se passe après l'ajout d'une coiffe (CAP) et d'une queue polyA.*

Les quantités d'ARNm, provenant de la transcription des différents gènes, sont donc très variables en fonction des protéines requises par la cellule à un instant donné. Le transcriptome est l'ensemble des ARNm présents dans la cellule à un instant donné. Sa connaissance représente une information de premier ordre pour la compréhension du fonctionnement de la cellule et pour la détermination des mécanismes qu'elle met en place dans des conditions particulières. C'est avec l'idée d'avoir une image du transcriptome que les puces à ADN ont été conçues.

Les puces à ADN sont constituées d'une surface plane sur laquelle sont fixées des sondes, qui sont des séquences oligonucléotidiques capables de se lier aux transcrits (les molécules d'ARNm) d'un organisme. Les transcrits marqués, d'un échantillon biologique, sont incubés avec la puce à ADN, permettant ainsi leur hybridation aux sondes de capture leur correspondant. Ceci est possible puisque l'ARNm résultant de la transcription d'un gène en a la séquence anti-complémentaire. La transcription se basant sur l'un des deux brins de l'ADN pour synthétiser un brin d'ARN anti-complémentaire, celui-ci garde la capacité de se lier à un brin d'ADN ayant la séquence du brin qui a permis sa synthèse. Une fois les transcrits hybridés aux sondes de la puce, les marqueurs sont excités et le signal émis est lu. Les marqueurs sont en grande majorité des fluorochromes qui sont capables d'absorber des photons à une longueur d'onde précise et d'en réémettre à une longueur d'onde supérieure.

3.1.2. Historique

La capacité qu'ont certaines molécules de se lier spécifiquement entre elles était déjà exploitée depuis les années 1960 avec, par exemple, des supports solides sur lesquels étaient présents des anticorps auxquels venaient se lier des antigènes spécifiques. Cette technique était, alors, utilisée pour le diagnostic de certaines maladies, par exemple. L'idée d'appliquer ce principe de liaison spécifique à l'ADN naquit dans le courant des années 1980 [136]. En effet, l'ADN étant composé de deux brins anti-complémentaires, chaque brin a la capacité de reconnaître très spécifiquement une séquence d'oligonucléotides qui lui est anti-complémentaire.

En 1991, Stephen Fodor et ses collaborateurs publient un article dans la revue Science [137]. Celui-ci traite de la synthèse, sur support physique, de certains composés biochimiques. La photolithographie était déjà utilisée en électronique où une photorésine, appliquée à la surface d'un substrat, est exposée à une radiation lumineuse [138]. En utilisant, lors de cette exposition, un masque ayant un certain motif, on reproduit celui-ci dans la photorésine. Cette technique permettait, déjà à l'époque, la fabrication simultanée d'un grand nombre de circuits électriques. Fodor et ses collaborateurs avaient su tirer parti de cette technique pour produire des polypeptides et des oligonucléotides en un minimum d'étapes. Le principe est de fixer, à la surface d'un support solide, une molécule portant un groupe protecteur photolabile. En exposant cette surface à la lumière, le groupe protecteur est éliminé, rendant la molécule accessible à d'autres molécules (par exemple des acides aminés pour la synthèse de polypeptides ou des nucléotides pour la synthèse d'oligonucléotides) portant elles-mêmes des

groupements protecteurs photolabiles empêchant la liaison d'autres molécules. En appliquant des masques lors des étapes d'exposition à la lumière, on choisit quelles molécules seront accessibles à d'autres molécules lors de l'étape suivante (Figure 14).

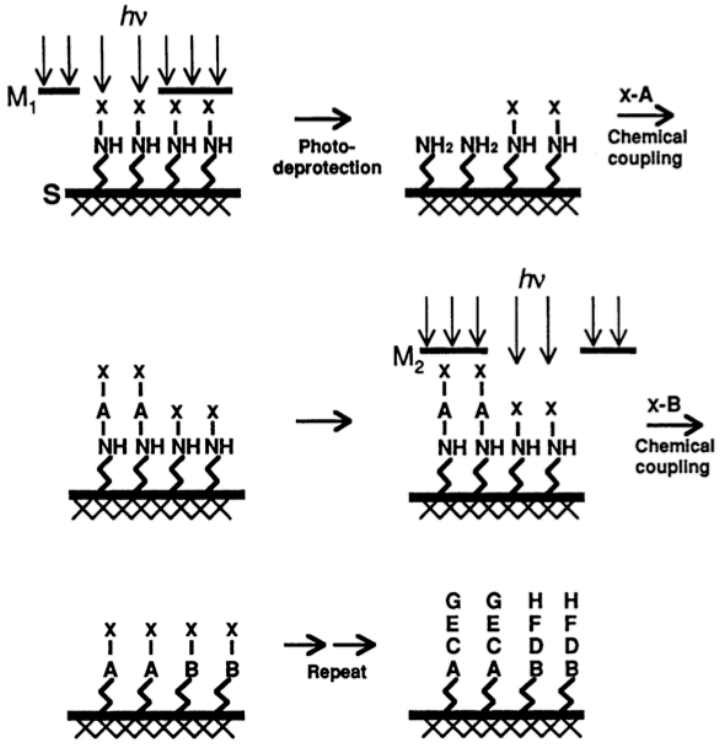


Figure 14. *Schéma représentant un substrat S auquel sont fixés des acides aminés dont la fonction amine est protégée par un groupement photolabile X. L'illumination à travers un masque M_1 mène à la dégradation de certains groupements X. Les acides aminés rendus libres du groupement X sont, dès lors, accessibles au composé A qui contient, lui aussi, le groupement protecteur photolabile X. Une nouvelle exposition à des radiations lumineuses en présence du masque M_2 permet d'éliminer les groupements X d'une autre région de la plaquette et de lier le composé B aux acides aminés nouvellement accessibles. En répétant plusieurs fois ce cycle, on peut synthétiser sur un support un ensemble de produits précis (source : [137]).*

Fodor et ses collaborateurs ont ainsi conçu plusieurs puces. Pour démontrer la rapidité de la technique, une puce de 1.024 polypeptides différents, composés de 0 à 10 acides aminés, a été produite en 10 étapes, montrant ainsi qu'en n étapes, l'on pouvait synthétiser 2^n composés différents. Pour les puces nucléotidiques, à l'époque, seuls 85 à 95% des liaisons entre les nucléotides des sondes étaient effectivement réalisées lors de la synthèse des puces. Cela était dû à la diffraction, la réflexion et la diffusion d'une partie de la lumière, lors des

phases d'exposition, qui entraînaient une élimination incomplète des groupements protecteurs et donc une liaison incomplète des composés. Cependant, la synthèse *in situ* par photolithographie présentait de nombreux avantages par rapport aux méthodes concurrentes (synthèse par aiguilles, approches utilisant l'ADN recombinant, …). Par exemple, 250.000 composés différents pouvaient, à l'époque, être synthétisés par centimètre carré et l'on pensait que ce nombre pourrait augmenter jusque dix milliards. Fodor et ses collaborateurs proposaient, dans leur article, plusieurs applications possibles de leur méthode parmi lesquelles se trouvait déjà la liaison des transcrits aux séquences d'ADN qui leur étaient anti-complémentaires.

L'ADNc (ADN complémentaire) est l'ADN retranscrit à partir de l'ARNm épissé mature. Sur les puces à ADN, destinées à révéler la quantité d'ARNm mature, il convient de fixer l'ADNc qui lui sera complémentaire. Bien qu'il existe plusieurs méthodes pour produire de l'ADNc, la plus commune est l'utilisation d'une enzyme appelée la transcriptase inverse qui, à partir d'un brin d'ARN, génère le brin d'ADN complémentaire [139].

En 1995, la revue Science publie un nouvel article traitant des puces à ADN. Cette fois, les auteurs sont Mark Schena et ses collaborateurs. Ils sont les premiers à utiliser cette technique pour mettre en évidence une différence d'abondance de transcrits entre plusieurs conditions expérimentales [133]. L'organisme modèle choisi fut la plante *Arabidopsis thaliana* car c'était, à l'époque, l'eucaryote supérieur possédant le plus petit génome connu. Quarante-cinq sondes, d'environ 1.000 bases chacune, représentant 45 gènes d'*Arabidopsis* et trois sondes représentant des gènes d'autres organismes, servant de contrôles négatifs, ont été déposées sur des lames de microscope. Ces puces à ADN ont permis de réaliser plusieurs expériences comparant les niveaux d'expression des gènes dans différentes conditions. Ces résultats ont ensuite été confirmés par des analyses par northern blot.

En 1997, la revue Science publie, à nouveau, un article clé dans l'histoire des puces à ADN [140]. Joseph DeRisi et l'équipe de Patrick Brown, qui était déjà impliqué dans l'article de Mark Schena, sont les premiers à proposer une étude utilisant des puces à ADN où un génome entier est représenté, en l'occurrence celui de la levure *Saccharomyces cerevisiae.* Cet organisme était particulièrement attractif, à l'époque, pour réaliser ce projet puisque ses 6.400 gènes étaient connus, facilement repérables au sein de son ADN et qu'il existait déjà un grand nombre d'outils pour leur analyse. La construction de puces, où 6.400 sondes différentes représentaient le génome de *Saccharomyces cerevisiae*, permit aux auteurs de réaliser trois expériences distinctes afin de savoir quand les gènes étaient transcrits.

La première expérience réalisée par DeRisi et ses collaborateurs permit de mettre en évidence les gènes qui étaient régulés, au niveau transcriptionnel, lors du passage du métabolisme anaérobique au métabolisme aérobique. Lors de la seconde et de la troisième expérience, les auteurs ont pu mettre en évidence les gènes régulés au niveau transcriptionnel chez des levures dont le gène codant pour le co-répresseur de transcription TUP1 était délété et chez des levures qui surexprimaient l'activateur de transcription YAP1.

En février 2001, le séquençage brut du génome humain est publié. Celui-ci fait, en fait, l'objet de deux publications [141, 142] suite à la « course au génome humain » qui s'était déclarée trois ans auparavant entre le consortium international public et la société privée Celera Genomics, dirigée par Craig Venter. Ce séquençage est, à l'époque, encore imparfait : certaines parties sont manquantes et d'autres sont erronées. C'est seulement en 2004 que le consortium international public publie, dans la revue Nature, la séquence complète du génome humain [143].

Depuis 2004, plusieurs sociétés commerciales proposent des puces à ADN permettant l'analyse de l'expression de l'ensemble des gènes du génome humain [144]. Ceci a plusieurs avantages : un coût réduit par rapport à plusieurs puces où sont représentées les différentes parties du génome, une diminution du nombre d'étapes dans la préparation des puces et une variabilité expérimentale réduite due à cette diminution. Différentes techniques sont employées par ces sociétés pour produire les puces : Agilent synthétise les oligonucléotides *in situ* par jets d'encre (c'est-à-dire en déposant les bases les unes après les autres sur des lames de verre), Affymetrix utilise la photolithographie avec des masques pour la synthèse *in situ*, Nimblegen fait de même, mais en utilisant des micromiroirs, enfin, Applied Biosystems synthétise les sondes oligonucléotidiques à part, avant de les déposer sur la puce. À l'époque, Agilent commercialisait des puces où 41.000 transcrits étaient représentés, Affymetrix, 47.000, Nimblegen, 38.000, et Applied Biosystems, 30.000.

Depuis la publication de la séquence du génome humain, les connaissances s'y rapportant ne cessent d'augmenter. De plus en plus d'études montrent que la réalité biologique est bien plus complexe que le dogme central de la biologie moléculaire. Par exemple, la plupart des transcrits se trouvant, à un moment donné, dans une cellule eucaryote sont le produit de l'épissage alternatif (les différentes combinaisons possibles d'exons) d'un nombre plus restreint de gènes [145]. Ces dernières années, d'autres niveaux de régulation génique ont été découverts. Citons, notamment, les différents mécanismes de régulation post-transcriptionnelle par les miRNA capables de se lier aux ARNm afin d'empêcher leur

traduction en protéines [146] ou de permettre leur dégradation [147] ou même de se lier aux gènes afin de les rendre silencieux par méthylation [148] (il s'agit alors d'une régulation transcriptionnelle), ou encore la régulation post-traductionnelle par phosphorylation des protéines, rendant celles-ci actives ou inactives [149]. Tous ces phénomènes rendent le défi de la compréhension du fonctionnement cellulaire plus attractif que jamais, les puces à ADN tentent donc d'évoluer avec ces découvertes. C'est ainsi que l'on a assisté à l'émergence des « Exon Arrays » dont les sondes sont censées représenter tous les exons du génome. Ceci est rendu possible par la capacité à synthétiser des oligonucléotides sur des surfaces de plus en plus petites.

Parallèlement à l'évolution des puces à ADN, on a également vu naître sur Internet des bases de données rendant publiques les données brutes provenant des expériences réalisées avec des puces à ADN. Les meilleurs exemples de telles bases de données sont Gene Expression Omnibus (GEO) et ArrayExpress [150]. Ces bases de données permettent le dépôt et l'archivage des données génomiques générées par des techniques à haut débit, et soumises par la communauté scientifique, sous forme numérique. Ces données sont ensuite consultables et téléchargeables publiquement. La technique des puces à ADN devenant de plus en plus courante, un format standard de présentation des données, le MIAME [151], s'est imposé au fil du temps et s'est même vu imposé par les revues scientifiques pour qu'un article relatant l'utilisation de puces à ADN soit publié. Il est important de noter que, grâce à la possibilité de numériser de grandes quantités de données, c'est la première fois dans l'histoire de la biologie qu'une telle masse d'informations est disponible à tout un chacun pour procéder à une ré-analyse de résultats. On voit ainsi un même ensemble de données générer plusieurs résultats par les différentes façons de les analyser.

3.1.3. Principaux types de puces à ADN

Il existe, aujourd'hui, trois grands types de puces à ADN [152] : les macroarrays, les microarrays et les puces à oligonucléotides (cette nomenclature peut varier selon les auteurs).

a) Les macroarrays

Pour la construction des macroarrays, des gouttelettes contenant des séquences d'ADN, généralement de 60 nucléotides, sont déposées sur une membrane de nylon. Ces séquences d'ADN constituent les sondes du macroarray. Il peut y avoir jusqu'à 2.400 gouttelettes déposées sur une membrane, chacune étant constituée d'un grand nombre de copies d'une séquence d'ADN. Il peut donc y avoir jusqu'à 2.400 sondes différentes sur un

macroarray. Une fois déposées sur la membrane, les gouttelettes ont une taille comprise entre 0,5 et 1 mm. L'échantillon, à tester par ce type de puce à ADN, doit être de l'ADN marqué au ^{32}P (un isotope du phosphore). Pour ce faire, l'ARN de l'échantillon est rétro-transcrit en présence de ^{32}P [153]. Une seule condition expérimentale peut être testée par puce. Cela implique que pour mesurer une expression différentielle des gènes entre deux conditions, il faudra utiliser au moins deux macroarrays. Une fois la détection de la radioactivité du macroarray réalisée, l'on obtient une image où les points (ou spots), correspondant aux différentes sondes de la puce, sont plus ou moins intenses. Plus un point est intense, plus l'ADN, rétro-transcrit de l'ARN de l'échantillon, a pu s'hybrider, et donc, plus le gène représenté par la sonde était exprimé dans l'échantillon (Figure 15).

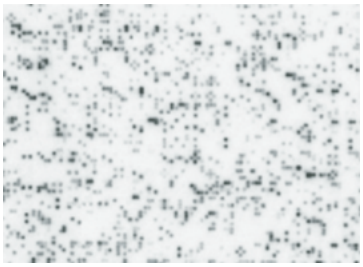


Figure 15. *Résultat de la lecture de la radioactivité d'un macroarray (source : [154]).*

b) Les microarrays

Les microarrays ont été développés à Stanford et sont illustrés dans la section précédente par l'article de Joseph DeRisi [95]. Lors de leur fabrication, des gouttelettes, contenant un grand nombre de copies d'une sonde (séquence de 30 à 70 nucléotides), sont déposées sur une lame de verre traitée par un revêtement chimique qui permet de fixer l'ADN. Jusqu'à 10.000 gouttelettes peuvent être déposées par centimètre carré de support, le diamètre de chaque dépôt mesurant environ 100 µm. Ce type de puce à ADN demande de l'ADN comme échantillon, celui-ci doit être marqué à la cyanine 3 pour une condition (par exemple, un traitement avec un médicament) et à la cyanine 5 pour une seconde condition (par exemple, une condition contrôle). L'ADN marqué à la cyanine 3 et l'ADN marqué à la cyanine 5 sont ensuite placés ensemble sur la puce. L'un des avantages de cette technique est donc de pouvoir tester deux conditions sur une même puce. En fonction de la lumière émise après excitation, l'on peut déterminer dans quelle condition un transcrit était présent en plus grande quantité (Figure 16).

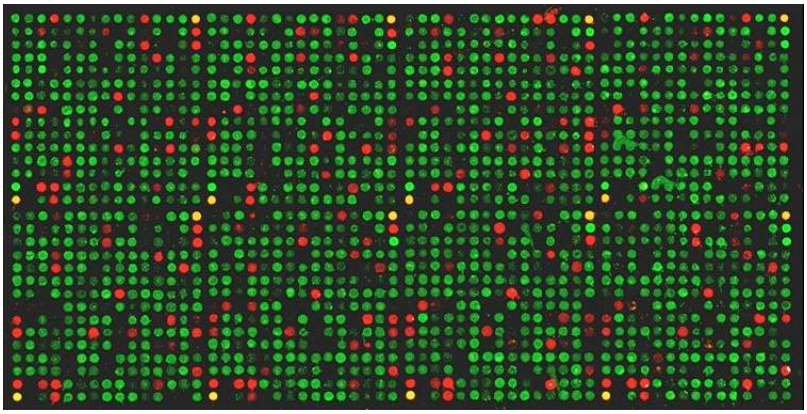


Figure 16. *Résultat de la lecture de la fluorescence d'un microarray (source : [155]).*

c) Les puces à oligonucléotides

Les puces à oligonucléotides sont des lames de verre à revêtement chimique où des sondes de 25 à 70 nucléotides sont synthétisées *in situ* par photolithographie [156] ou par impression de type « jet d'encre » [157]. Elles se distinguent des autres types de puces notamment par le fait qu'elles présentent la plus haute densité de spots, puisqu'il peut y en avoir de l'ordre du million par centimètre carré de support, chacun ayant une taille de moins de 20 µm. Les puces à oligonucléotides de la marque Affymetrix sont les plus répandues et nécessitent de l'ARN marqué à la biotine comme échantillon. Une révélation à la streptavidine, elle-même couplée à un fluorochrome, s'en suit. Contrairement aux microarrays présentés dans la section précédente, une seule condition peut être analysée par puce. En effet, avec ce type de puce, la couleur est toujours la même, seule son intensité diffère en fonction du niveau d'expression des gènes (Figure 17).

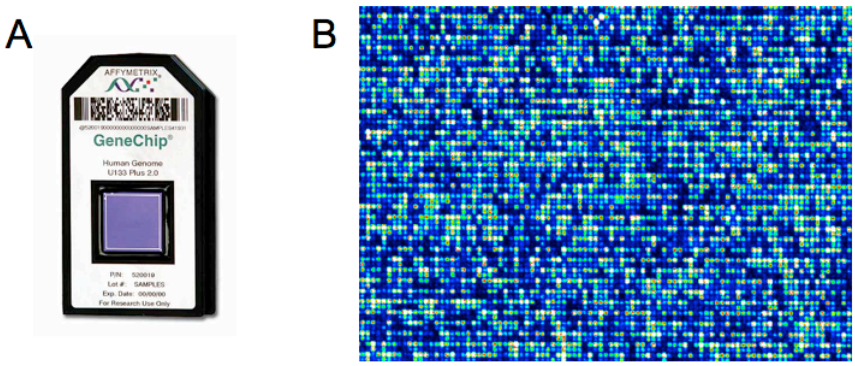


Figure 17. A) *Puce à oligonucléotides de la marque Affymetrix.* ***B)*** *Résultat de la lecture de la fluorescence d'une puce à oligonucléotides (source : www.affymetrix.com).*

3.1.4. Applications

a) La détermination du profil d'expression des gènes

Il s'agit de l'utilisation la plus courante des puces à ADN. En hybridant un échantillon sur une puce à ADN, on obtient une mesure de l'abondance des ARNm présents dans l'échantillon. En effectuant cette mesure dans différentes conditions ou sur différents types cellulaires, on peut identifier les gènes co-régulés dans certaines réponses cellulaires spécifiques, prédire la fonction de gènes non caractérisés car ceux-ci sont co-régulés avec des gènes de fonction connue, ou encore mettre en évidence des réseaux de régulation de voies biochimiques complexes [158].

Les puces à ADN sont aussi des outils de diagnostic. Notamment, elles permettent une nouvelle classification plus fine des types de cancers ou d'identifier de nouveaux gènes cibles de substances thérapeutiques, ainsi que la réponse cellulaire à un traitement [159].

b) L'hybridation génomique comparative

La génomique comparative permet de déterminer les variations du nombre de copies des gènes se trouvant dans l'ADN. En effet, chez l'être humain, hormis dans les cellules germinales et pour certains gènes se trouvant sur les chromosomes sexuels, tous les gènes sont en deux copies, l'une provenant de la mère, l'autre du père. Cependant, dans certaines pathologies dont le cancer, ce nombre se voit parfois modifié : des séquences d'ADN génomique peuvent être amplifiées [160] ou délétées. En mettant l'ADN de tissu normal marqué avec un fluorochrome (généralement la rhodamine, qui émet dans le rouge) avec l'ADN du tissu à tester marqué avec un autre fluorochrome (généralement la fluorescéine, qui émet dans le vert) sur une puce à ADN, on peut mesurer la variation du nombre de copies des gènes. En effet, si lors de la détection, le signal est rouge, alors il y a délétion dans le tissu testé puisque l'ADN du tissu normal s'est hybridé en plus grande quantité. Par contre, si le signal est vert, c'est l'ADN du tissu testé qui était en plus grande quantité, il y a donc eu amplification. Enfin, si le signal est jaune, alors les quantités d'ADN pour ce gène particulier étaient les mêmes dans les deux tissus.

c) L'identification d'organismes

Il existe des puces à ADN composées de séquences représentant des gènes spécifiques de certains organismes. En utilisant ces puces à ADN particulières, on peut détecter la

présence de certains micro-organismes dans l'alimentation, de champignons dans des cultures cellulaires ou encore de pathogènes dans le cas des maladies [161].

d) L'immuno-précipitation de la chromatine sur *chip*

Plus connue sous le nom de « ChIP on chip » pour « Chromatin ImmunoPrecipitation on chip », il s'agit de la combinaison de deux techniques pour caractériser les séquences d'ADN auxquelles peuvent se lier certaines protéines [162]. L'application la plus courante étant la détermination de la séquence à laquelle se lie un facteur de transcription. Après cross-linking, l'ADN est extrait des cellules avec les protéines qui lui sont liées. Cet ADN est ensuite fragmenté ; à ce moment, les protéines lui sont toujours liées. Cet ensemble de fragments d'ADN, dont certains sont liés à une protéine, sont mis en présence d'un anticorps qui reconnaît et lie spécifiquement la protéine d'intérêt (par exemple, le facteur de transcription étudié). Les complexes ADN-protéine-anticorps ainsi formés sont précipités et séparés du reste. Ensuite, l'ADN des complexes est séparé de ceux-ci (c'est la fin de la partie immunoprécipitation) pour être placé sur une puce à ADN afin de déterminer à quelle séquence anti-complémentaire il se lie et ainsi en déduire sa séquence propre et donc la séquence du site de liaison de la protéine étudiée.

e) La détection de Single Nucleotide Polymorphism (SNP)

Bien que tous les individus d'une espèce possèdent les mêmes gènes, ceux-ci présentent, néanmoins, des variations rendant chaque individu unique. Ces variations sont dues en grande partie à des SNP, c'est-à-dire des différences, entre deux allèles, d'un seul nucléotide dans une séquence d'ADN précise [163]. Ces différences peuvent toucher des parties codantes ou non-codantes de l'ADN. Dans le cas de parties codantes, la différence peut éventuellement entraîner un changement dans la séquence d'acides aminés de la protéine résultant de la transcription et de la traduction de cette séquence d'ADN.

Les SNP Arrays sont des puces à ADN où se trouvent des séquences oligonucléotidiques correspondant aux différentes séquences possibles de gènes. En plaçant l'ADN d'un individu sur de telles puces, on peut détecter les SNP qui caractérisent ses gènes puisque cet ADN se liera spécifiquement aux séquences qui lui sont anti-complémentaires sur la puce.

f) La détection d'épissage alternatif

Comme cela a déjà été vu précédemment, les gènes eucaryotes sont composés d'introns et d'exons. Seuls les exons sont transcrits et traduits en protéines. Les différentes

combinaisons d'exons d'un gène donnent alors des transcrits, et *in fine* des protéines, différents. Ce processus d'épissage peut être étudié à l'aide de puces à ADN de deux manières différentes. La première consiste à utiliser des « exon junction arrays ». Les sondes de ceux-ci représentent les jonctions potentielles entre exons. Ainsi, le signal de ces puces indique quelles combinaisons d'exons sont observées dans l'échantillon [164]. La seconde technique permettant l'étude de l'épissage alternatif sont les « exon arrays ». Sur ces puces particulières est représenté l'ensemble des exons du génome. Ceci a été, récemment, rendu possible grâce à la densité de synthèse des sondes toujours plus élevée. Grâce à une étude combinatoire du signal de ces puces, on peut connaître de quelle manière les gènes ont été épissés [165].

g) La détection de fusions de gènes

Il arrive, notamment dans certaines pathologies, que des gènes fusionnent, donnant naissance à des transcrits nouveaux [166]. Certaines puces à ADN sont conçues pour détecter ces transcrits anormaux. En effet, les sondes qui les composent représentent des fusions de transcrits. Il est important de noter que ce type de puces à ADN sont rarement des puces commerciales produites à grande échelle, il s'agit, le plus souvent, de puces conçues par des équipes pour étudier un phénomène particulier et produites en très faible quantité.

3.2. L'analyse des résultats issus de puces à ADN

3.2.1. Généralités

Comme nous venons de le voir, la diversité de types de puces à ADN et de leurs applications est très grande. Il existe donc également une grande diversité de types d'analyse. Ce travail ne s'intéressant qu'à l'expression différentielle de gènes entre deux conditions, et n'utilisant que des puces à oligonucléotides de la marque Affymetrix, seule l'analyse selon ces paramètres sera envisagée. Ceci permettra d'éviter de détailler tous les types d'analyse de puces à ADN qui existent, ce qui n'est nullement le but du présent manuscrit.

3.2.2. Particularités de la technologie Affymetrix

Les GeneChips Affymetrix sont les puces à ADN les plus répandues, tant dans les chiffres de ventes que dans les bases de données telles que GEO ou ArrayExpress. Les sondes y sont synthétisées *in situ* par photolithographie. En règle générale, l'analyse de l'abondance du transcrit de chaque gène est rendue possible par la présence de 11 à 20 sondes « perfect match » de 25 nucléotides et par le même nombre de sondes « mismatch » ne différant que

par le nucléotide central afin de mesurer l'hybridation non spécifique [167]. L'ensemble de ces sondes forme un « probe set ». Les différentes sondes d'un probe set ne sont pas synthétisées dans des zones contiguës pour des raisons d'échantillonnage des conditions expérimentales, mais elles sont dispersées sur la puce, qui porte des repères permettant de retrouver sur l'image la localisation précise de chaque sonde. Un probe set identifie donc une séquence d'ADN transcrite, mais par abus de langage on parlera souvent de gène (un probe set correspond à un « gène »).

Les sondes et les conditions d'hybridation sont conçues de manière à maximiser leur sensibilité, c'est-à-dire leur capacité à réellement s'hybrider avec la cible, leur spécificité, c'est-à-dire leur capacité à ne s'hybrider qu'à une seule cible, ainsi que leur reproductibilité, permettant ainsi de faire la différence entre le « bruit » et les signaux d'intérêt, ainsi qu'entre deux séquences cibles de structures très proches. Ces qualités sont testées expérimentalement pour certaines sondes, mais sont modélisées pour la plupart d'entre elles, qui sont donc de qualité variable.

3.2.3. Obtention et numérisation de l'image

Chaque sonde occupe une cellule de 24 µm sur 24 µm sur la puce (ce chiffre évoluant constamment, il peut ne plus être d'actualité au moment de la lecture de ce document). Cet espace est converti en 64 pixels (8 x 8 pixels, ce chiffre peut également être obsolète) (Figure 18). L'intensité de la cellule est calculée à partir des pixels centraux, les pixels périphériques étant ignorés pour éviter des effets de bords ou d'autres artéfacts [167]. À l'aide d'un logiciel d'analyse d'image, le niveau de fluorescence de l'ARNm marqué se liant à chacune des sondes sur la puce à ADN est déterminé [168].

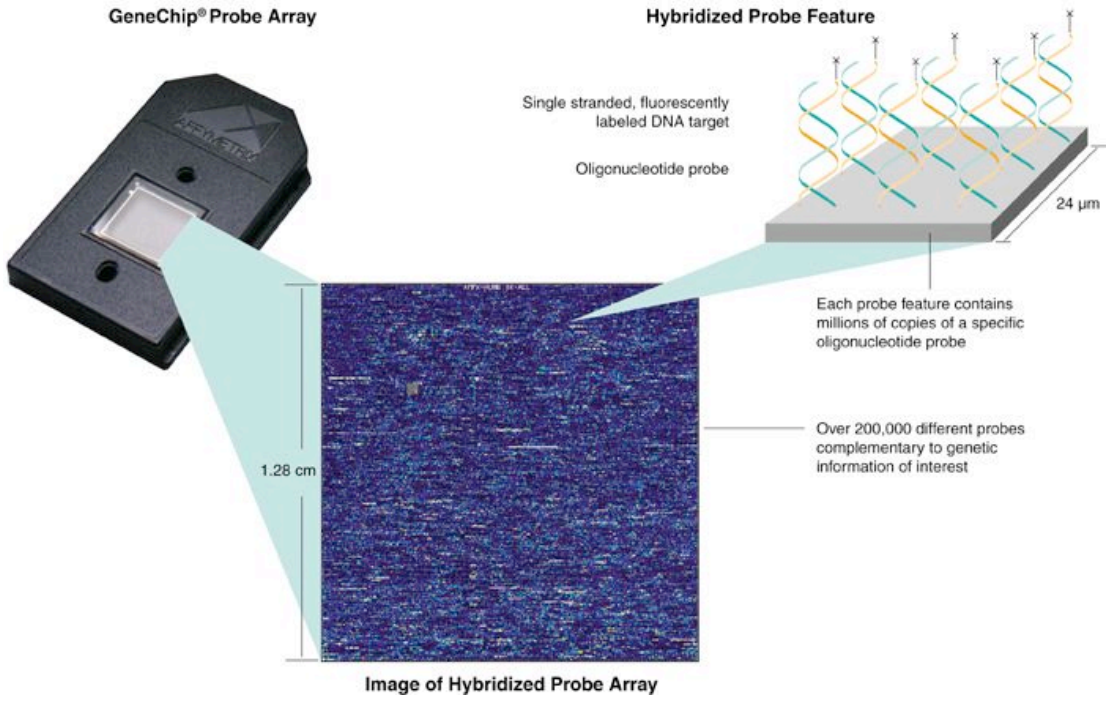


Figure 18. *Détail d'une cellule de 24 µm sur 24 µm d'un GeneChip Affymetrix (source : www.affymetrix.com).*

Un autre logiciel numérise, ensuite, l'image de telle sorte que les signaux lumineux les plus intenses, correspondant aux sondes auxquelles l'échantillon s'est le plus hybridé, ont les valeurs numériques les plus élevées. Ces données numériques sont stockées, dans le cas des GeneChips Affymetrix, dans un fichier au format .CEL.

3.2.4. Obtention d'un affybatch

Un affybatch est un terme spécifique de la technologie Affymetrix. Il s'agit des données numériques, contenues dans le fichier CEL, rapportées aux identifiants des gènes. Le fichier CEL n'est qu'une succession de nombres inutilisables en l'état. Pour y donner un sens, chaque nombre, correspondant à l'intensité du signal d'une sonde précise, doit être lié à un identifiant de gène grâce à un Chip Definition File (CDF), fichier mis au point par Affymetrix et qui relie à un gène donné les 11 à 20 sondes perfect match et les 11 à 20 sondes mismatch d'un « probe set ». Les sondes représentant les gènes reflètent l'état des bases de données génomiques au moment de leur conception, il y a plusieurs années. Depuis lors, l'information génomique, et donc les arguments pour attribuer des sondes données à un gène, ont évolué et des « CDFs alternatifs » ont vu le jour.

En 2004, après qu'Affymetrix ait mis à la disposition du public les séquences de ses sondes, Laurent Gautier et ses collaborateurs publient un article dans la revue BMC

Bioinformatics qui, pour la première fois, propose une comparaison des séquences des sondes du GeneChip HG-U133A (l'une des puces Affymetrix représentant l'entièreté du génome humain) à celles que l'on peut trouver dans les bases de données [169]. En effet, les annotations des génomes séquencés avaient rapidement progressé depuis la conception et la commercialisation des puces par Affymetrix. La première chose que les auteurs ont constatée est qu'il existe des sondes qui correspondent à plusieurs transcrits dans la base de données RefSeq (la base de données de NCBI concernant la séquence des transcrits). Ces transcrits n'ayant, parfois, aucun rapport entre eux. Ensuite, ils ont montré que 8% des « probe sets » étaient redondants. En effet, parfois jusqu'à six « probe sets » correspondaient au même transcrit. De plus, les auteurs ont constaté que le nombre de sondes par « probe set » n'avait pas besoin d'être aussi élevé que ce qu'Affymetrix proposait. Finalement, Gautier et ses collaborateurs ont montré que seulement 36% des probe sets conçus par Affymetrix correspondaient à un transcrit unique répertorié dans la base de données RefSeq. C'est pour toutes ces raisons que cette équipe proposait un CDF alternatif dont les probe sets étaient constitués de sondes ne correspondant qu'à un seul transcrit selon RefSeq.

À la suite de cet article, plusieurs autres équipes se sont lancées dans la conception de CDFs alternatifs [170]. Toutes utilisaient des stratégies plus ou moins différentes, ainsi que les références de bases de données diverses. Mais toutes arrivaient à la même conclusion : il était nécessaire de mettre à jour les CDFs fournis par Affymetrix, sans quoi l'interprétation des résultats était compromise. Certains auteurs allaient même, avec raison, jusqu'à conseiller de refaire, en utilisant les CDFs alternatifs, toutes les analyses ayant déjà été réalisées.

En 2007, Hongfang Liu et ses collaborateurs publient un article dans la revue Bioinformatics [171]. Celui-ci propose une base de données, accessible au public, reprenant, pour la première fois, des CDFs alternatifs pour tous les GeneChips commercialisés par Affymetrix jusqu'alors. Ces CDFs regroupent les sondes en probe sets selon l'information venant de RefSeq et GenBank (une base de données concernant la séquence des gènes).

3.2.5. Prétraitement des données

Les étapes de prétraitement consistent à essayer de rendre les données numériques brutes les plus représentatives possibles de l'information biologique. Le prétraitement se fait classiquement en trois (ou quatre) étapes : la correction du bruit de fond, la normalisation, la correction PM/MM et la « summarization ».

Une caractéristique de la technologie Affymetrix, importante pour comprendre certaines étapes du prétraitement, est la synthèse, pour chaque sonde, d'une séquence exacte (Perfect Match ou PM) et d'une séquence dont une base diffère (la treizième) par rapport à la séquence du gène qu'elle représente (MisMatch ou MM). La sonde mismatch sert à mesurer l'hybridation non spécifique, c'est-à-dire des transcrits qui ne sont pas la cible de la sonde, et sa valeur doit être soustraite de la valeur de la sonde perfect match.

Sur une puce, l'intensité mesurée est une combinaison d'image de fond et d'image résultant d'hybridations. Cependant, une partie des hybridations n'est pas spécifique. De plus, l'intensité du signal est affectée par du bruit optique provenant du scanner. L'existence du bruit de fond rend souvent l'estimation du niveau exact d'hybridation très difficile. Différentes méthodes ont été proposées pour cette étape (Figure 19) ; elles peuvent être rangées en trois catégories, suivant qu'elles utilisent :

- la moyenne ou la médiane des valeurs d'intensités des sondes comme estimateur du bruit de fond global [172],
- les pixels se trouvant près du spot pour estimer le bruit de fond local,
- des filtres non-linéaires utilisant la moyenne pondérée des valeurs des pixels voisins pour remplacer la valeur d'un pixel aberrant [173].

Figure 19. *Représentation schématique de l'effet de la correction du bruit de fond.*

La normalisation sert à équilibrer les niveaux d'expression entre les régions d'une même puce, et entre puces d'une même expérience, de manière à pouvoir comparer ce qui est comparable (Figure 20). Le statisticien, habitué à l'expression « normalisation » en référence à une distribution Gaussienne, notera que dans ce contexte, normalisation signifie « homogénéisation » ou « standardisation » mais ne fait aucune référence à la distribution « normale » (Gaussienne). Il existe plusieurs méthodes de normalisation linéaires ou non [174] (qui ne seront toutefois pas détaillées dans la présente étude, car elles n'en sont pas l'objet).

Figure 20. *Représentation schématique de l'effet de la normalisation.*

Utiliser l'information contenue dans les sondes mismatch pour corriger l'information véhiculée par les sondes perfect match semblait une bonne idée *a priori*. Malheureusement, il semblerait que cette correction soit, elle-même, source d'erreurs. En effet, Irizarry et ses collègues ont montré sur un jeu de données qu'un tiers des sondes MM donnait un signal d'intensité supérieure à celui de la sonde PM. Cela montrait d'une part que les sondes ne correspondaient pas toujours aux transcrits qu'elles étaient censées cibler (d'où l'intérêt des CDF alternatifs), et d'autre part que le changement du nucléotide central dans la sonde mismatch ne permettait pas de mesurer le bruit dû aux hybridations non spécifiques. Comme, lors de la correction PM/MM, on soustrait l'intensité du signal de la sonde mismatch de celle de la sonde perfect match, cela entraîne des intensités négatives [175]. C'est pour cela que la majorité de la communauté scientifique n'effectue pas cette correction. Par contre, il est intéressant de noter que les CDFs alternatifs se servent des sondes mismatch dans la définition des probe sets quand leur séquence correspond effectivement à un transcrit.

Après prétraitement des sondes, la summarization combine toutes les valeurs d'intensité des sondes relatives à un probe set en une seule valeur d'expression [175].

Il existe une grande variété de méthodes de prétraitement qui combinent les différentes variantes des étapes qui viennent d'être expliquées [176]. Ces méthodes sont implémentées dans des algorithmes qui génèrent des résultats différents. Des erreurs introduites au niveau d'une étape de prétraitement altéreront les traitements ultérieurs [173]. Les performances des méthodes statistiques qui se trouveront en aval des données prétraitées dépendront donc fortement des options choisies pour le prétraitement. Dans les faits, c'est souvent la combinaison d'un prétraitement et d'un traitement donnés qui atteint une performance intéressante. L'étude d'une méthode statistique particulière est donc indissociable des étapes qui l'ont précédée, ce qui explique que nous nous y soyons attardé quelque peu.

Les dizaines de méthodes disponibles pour chaque étape de prétraitement et les combinaisons de ces méthodes peuvent potentiellement générer des milliers de façons de prétraiter les données. Parmi les plus célèbres, citons MAS 5.0, RMA et GCRMA [175, 177]. Bien que de nombreux points différencient ces trois algorithmes, nous allons brièvement décrire les principaux.

Historiquement, MAS 5.0 est le premier qui a pu être utilisé par les chercheurs. Et pour cause, MAS 5.0 est l'algorithme fourni par Affymetrix. MAS 5.0 procède aux quatre étapes décrites précédemment. Cependant, comme mentionné plus haut, il se trouve que, dans un nombre non négligeable de cas, la sonde perfect match donne un signal plus faible que son

homologue mismatch, conduisant à une valeur d'expression négative. C'est pour cette raison que Rafael Irizarry et ses collaborateurs ont proposé, en 2003, le logiciel RMA (Robust Multi-array Analysis) qui n'effectuait que trois étapes du prétraitement, évitant la correction PM/MM [175]. Un an plus tard, en 2004, Zhijin Wu, alors dans l'équipe d'Irizarry, propose une version évoluée de RMA : GCRMA (GeneChip Robust Multi-array Analysis) [177]. Cet algorithme, en plus d'éviter la correction PM/MM, applique un modèle statistique à la correction du bruit de fond qui prend en compte la séquence des sondes. En effet, ces auteurs ont constaté qu'une part du bruit de fond était due à l'hybridation non spécifique que permettaient plus facilement les sondes riches en guanine et en cytosine, puisque ces nucléotides sont capables de former trois ponts hydrogène, contrairement à l'adénine et à la thymine, qui ne peuvent en former que deux.

3.2.6. Traitement statistique des données

Des méthodes statistiques ont très vite été utilisées après l'apparition des puces à ADN pour extraire l'information biologique et pour estimer l'incertitude, causée par l'importance de la quantité et la variation intrinsèque des données obtenues sur une puce. Ces analyses se situent en aval du prétraitement des données brutes.

La première méthode pour identifier des gènes exprimés de manière différentielle a été le fold change [178]. Il consiste à évaluer le logarithme en base 10 du rapport d'expression de chaque gène entre deux conditions et de considérer les gènes qui diffèrent plus qu'un seuil arbitraire, comme étant exprimés de manière différentielle. Ce n'est pas un test statistique inférentiel (c'est-à-dire dont le risque d'erreur est contrôlé) puisqu'il ignore toute notion de variabilité des réplicats et ne dispose donc pas de valeurs qui mesure sa confiance. Il présente l'avantage d'être simple et applicable à des jeux de données pour lesquels une seule puce est disponible par condition.

Lorsque plusieurs réplicats sont disponibles par condition, la question peut être abordée d'un point de vue statistique [179]. En effet, dès lors que les conditions sont répétées sur plusieurs puces, la mesure de la variabilité devient possible et un test statistique donnant une estimation du risque d'erreur peut être effectué. Plusieurs méthodes statistiques ont récemment émergé pour répondre à cette demande. La plupart sont des variantes du test t de Student [180]. Le test t de Student classique a été le premier à être utilisé. Cependant, ce test est peu puissant car le nombre de réplicats pour chaque condition reste petit ; l'estimation de la variance est donc peu précise et le nombre de degrés de liberté faible. Cela veut dire que

seules les différences importantes peuvent être statistiquement significatives. Des versions plus sophistiquées du test t de Student, jouant sur l'estimation de la variance, sont donc apparues. En 2001, Baldi et Long ont proposé le « regularized t test » qui prend en compte, dans l'estimation de la variance de l'expression d'un gène, la variance de l'expression d'un certain nombre de gènes ayant une variance proche [181]. En 2008, Fabrice Berger proposait le « Window t test » [182], développé au sein de notre laboratoire. Celui-ci permet, entre autres améliorations par rapport au « regularized t test », de modifier le nombre de gènes dont la variance de l'expression servira au calcul de la variance de l'expression de chaque gène. En effet, le « regularized t test » impose un nombre fixé de gènes pour l'estimation de la variance de l'expression d'un gène. Cependant, plus le nombre de réplicats est grand, moins le nombre de gènes dont la variance de l'expression servira au calcul de la variance de l'expression d'un gène doit être grand. C'est pourquoi le « Window t test » permet de contrôler le nombre de gènes pour l'estimation de la variance de l'expression d'un gène.

Une expérience utilisant des puces à ADN donne donc typiquement une liste de gènes exprimés de manière différentielle. À partir de là, le biologiste va tenter d'interpréter cette liste en y cherchant des gènes impliqués dans les mêmes voies, partageant les mêmes annotations, ou se trouvant dans les mêmes régions chromosomiques.

Depuis ces dernières années, une nouvelle approche s'est développée : il s'agit non plus de faire les tests statistiques individuellement sur chacun des gènes représentés sur la puce à ADN, mais de les faire sur des groupes de gènes biologiquement liés, c'est-à-dire partageant une même fonction, appartenant à une même voie de signalisation ou participant au développement d'une même maladie, par exemple. Ces groupes de gènes peuvent être établis grâce à des banques de données telles que Gene Ontology [183] ou KEGG [184].

Cette méthode présente comme premier avantage de permettre aux chercheurs de tester directement un ou plusieurs groupes de gènes d'intérêt plutôt qu'un nombre énorme de gènes individuels. Ensuite, alors qu'une approche classique ne distingue pas les faibles « fold changes » à cause du grand nombre de gènes testés, de la grande variabilité entre les individus et du nombre limité d'échantillons, une approche par groupes de gènes permet de distinguer ces faibles « fold changes » grâce à un gain de puissance.

En 2004, Goeman et ses collègues ont développé le Global Test [185]. Ils l'ont appliqué au jeu de données de Golub [186] qui distingue la leucémie myéloïde aiguë et la leucémie lymphoïde aiguë sur base de l'expression des gènes. Ils l'ont également appliqué à un jeu de données distinguant des cellules ayant reçu un choc thermique et des cellules

contrôles [185]. Le Global Test permet de tester et de comparer des groupes de gènes de tailles différentes.

En 2005, Goeman et son équipe donnaient à leur Global Test une extension associant le profil d'expression d'un ou plusieurs groupes de gènes à un temps de survie d'un patient cancéreux [187]. En 2005 également, Mansmann et Meister [188] ont proposé l'ANCOVA comme alternative au Global Test de Goeman. L'ANCOVA a été testée sur des puces à ADN traitant de différents stades de cancer du colon, et le groupe de gènes testé était celui de la voie de transduction du signal via la protéine p53. Dans ce cas particulier, l'ANCOVA s'est montrée plus puissante que le Global Test.

Enfin, en 2010, Fabrice Berger proposait une version modifiée de l'ANOVA 2 pour tester des groupes de gènes, cette méthode s'appelle FAERI [189]. Appliqué à plusieurs jeux de données traitant de l'hypoxie, FAERI s'est montré capable de détecter plus de groupes de gènes liés à la réponse à l'hypoxie que les autres méthodes.

3.2.7. Évaluation des méthodes de traitement statistique

Au vu du nombre croissant de méthodes de traitement statistique des données venant de puces à ADN, il a très vite fallu des outils pour comparer leurs performances. Rapidement, les courbes ROC (Receiver Operating Characteristic) se sont imposées. Il s'agit de graphiques représentant le taux de vrais positifs sur un axe et le taux de faux positifs sur l'autre. Cependant, ce type de représentation se révèle, en fait, très peu discriminant dans le cadre de la comparaison de traitements statistiques de données de puces à ADN [190]. Certains auteurs ont alors proposé l'utilisation des courbes FDR [182].

Une courbe FDR (False Discovery Rate) est un outil d'analyse de la variation de la sensibilité et de la valeur prédictive positive d'un test selon différents seuils de discrimination (Figure 21). L'axe des abscisses représente 1 – la valeur prédictive positive, c'est-à-dire le taux de découvertes erronées (False Discovery Rate, FDR), et l'axe des ordonnées représente la sensibilité, c'est-à-dire le taux de vraies découvertes (True Positive Fraction, TPF). La courbe est construite empiriquement en calculant la sensibilité et la valeur prédictive positive d'un test à différents seuils de discrimination. Au début de la courbe (partie gauche), les seuils de discrimination sont faibles et seules quelques valeurs sont retenues. Si beaucoup de ces valeurs sont des vrais positifs, la courbe aura tendance à rester collée à l'axe des ordonnées, si par contre, beaucoup de ces valeurs sont des faux positifs, elle aura tendance à s'en écarter. En avançant sur la courbe, les seuils de discrimination augmentent, pour arriver à 1, où toutes

les valeurs sont reprises. Idéalement, pour tout seuil de discrimination, toutes les valeurs retenues sont des vrais positifs, la sensibilité et la valeur prédictive positive sont alors égales à 1, la courbe reste collée contre l'axe des ordonnées, en tout point de la courbe l'abscisse vaut 0 et la surface sous la courbe est égale à 0 (Figure 22). À l'inverse, si pour tout seuil de discrimination, toutes les valeurs retenues sont des faux positifs, la sensibilité et la valeur prédictive positive sont égales à 0, la courbe reste collée à l'axe des abscisses, en tout point de la courbe l'ordonnée vaut 0 et la surface sous la courbe est égale à 0 aussi.

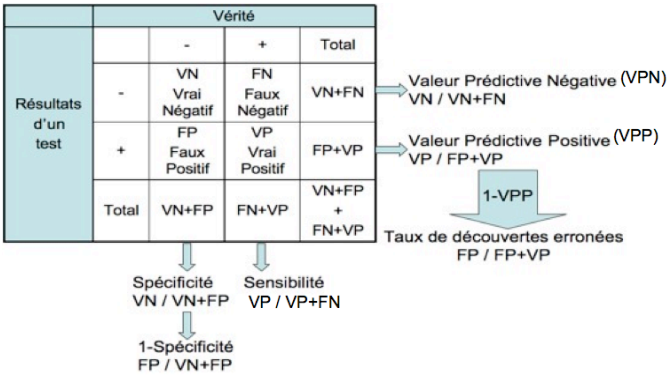


Figure 21. *Lors d'un test, les résultats sont positifs ou négatifs. De plus, ils peuvent être vrais ou faux. Le rapport des vrais positifs et de la somme des faux et des vrais positifs donne la valeur prédictive positive. En soustrayant cette valeur à 1, on obtient le taux de découvertes erronées (abscisses de la courbe FDR). Le rapport des vrais positifs et de la somme des faux négatifs et des vrais positifs donne la sensibilité (ordonnées de la courbe FDR).*

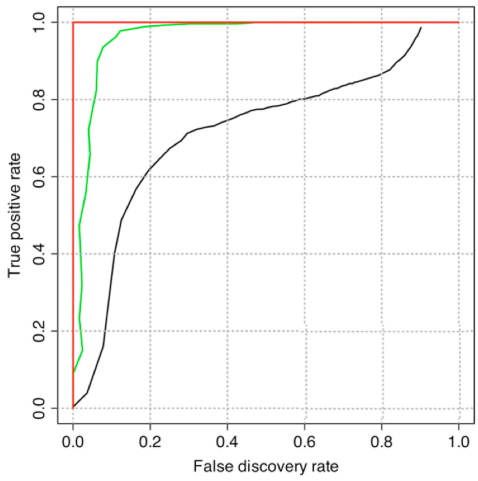


Figure 22. *Exemples de courbes FDR. L'axe des abscisses reprend le taux de découvertes erronées et l'axe des ordonnées la sensibilité. La courbe rouge représente le cas idéal. Dans cet exemple, la méthode représentée par la courbe verte est meilleure que la méthode représentée par la courbe noire. En effet, à sensibilité égale, le taux de découvertes erronées est plus important pour la méthode représentée par la courbe noire.*

Autant pour les courbes ROC que pour les courbes FDR, ces outils demandent, pour déterminer la spécificité et la sensibilité d'une méthode d'analyse, de connaître quels sont les gènes qui sont effectivement exprimés de manière différentielle ou non. Pour cela, plusieurs jeux particuliers de données existent. Les plus connus sont les carrés latins d'Affymetrix, LS 95 et LS 133 (2002), et le « golden spike » proposé par Choe et ses collègues en 2005 [191]. Dans ces jeux de données, la sous- ou surexpression des gènes est simulée en modifiant la concentration de l'ARN déposé sur la puce. Ainsi, les gènes qui doivent être détectés par le traitement statistique sont connus à l'avance.

En 2010, au sein de notre laboratoire, Benoît De Hertogh et Bertrand De Meulder proposaient une comparaison originale des méthodes de traitement statistique [192]. En effet, en construisant un jeu de données artificiel à partir de jeux de données se trouvant dans le domaine publique et réunissant un grand nombre de conditions expérimentales, ils fournissaient une meilleure approximation de la variabilité biologique que ne le faisaient les LS 95 et 133 et le « golden spike ». En rendant les gènes à détecter de plus en plus difficiles à détecter par une variabilité accrue et un bruit de fond croissant, ils ont montré que les méthodes de traitement les plus sensibles et spécifiques étaient le « shrinkage t test » [190], suivis par le « regularized t test » [181] et le « window t test » [182] lorsque le nombre de réplicats est de deux. Les performances absolues sont cependant limitées.

Avec la meilleure méthode, considérant une liste de gènes candidats dans laquelle 50% sont des faux positifs, 80% des vrais positifs sont retrouvés (20% de faux négatifs) avec n=3 si le fold change est de l'ordre de 4 et n=10 s'il est de l'ordre de 2. Considérant que dans certains mécanismes de régulation, de faibles variations de l'expression d'un gène peuvent parfois entraîner des conséquences métaboliques importantes, l'analyse d'un grand nombre de réplicats se justifie certainement. Cette approche est essentiellement permise par la métanalyse de plusieurs jeux de données rassemblés.

3.2.8. Corrections pour tests multiples

Le résultat de chacun des tests statistiques est exprimé en p value, qui représente l'erreur de type I, interprétée comme la probabilité d'observer par hasard un fold change significatif, alors que l'expression du gène n'est pas modifiée (il s'agit alors d'un faux positif). Les gènes sont classés par ordre croissant de p value et les gènes dont la p value est en dessous d'un niveau déterminé arbitrairement peuvent être considérés comme significatifs.

Il faut noter que les listes de gènes significatifs obtenues par différentes méthodes sur le même jeu de données diffèrent substantiellement.

Par ailleurs, les gènes qui ne sont pas repris dans la liste de gènes significatifs peuvent être, en réalité, exprimés de manière différentielle. Il s'agit de faux négatifs, correspondant à l'erreur de type II, de probabilité inconnue et fonction de la puissance du test. Comme nous l'avons dit, les tests t ne sont pas très puissants et le taux de faux négatifs est très élevé.

Le problème principal vient, cependant, de la multiplication des tests. Si l'on réalise l'analyse d'une puce comportant 20.000 probe sets (qui est un ordre de grandeur fréquemment atteint avec les GeneChips Affymetrix), 20.000 tests sont produits, qui génèrent, au seuil de 5%, 1.000 faux positifs. À supposer que l'on recherche la modification de quelques dizaines de gènes par les conditions expérimentales, ceux-ci, même s'ils ont passé le seuil statistique, seront « perdus » dans les faux positifs.

L'une des approches pour solutionner le problème des tests multiples est de contrôler le FWER (de l'anglais Family-Wise Error Rate) qui est la probabilité d'accumuler un ou plusieurs faux positifs dans la série de tests statistiques réalisée sur les gènes. Ce contrôle est réalisé en rendant plus strictes les conditions des tests. La correction de Bonferroni est la procédure la plus simple [193] : elle propose de diviser la valeur sous laquelle une p value doit se trouver pour considérer le gène comme significatif, par le nombre de tests, c'est-à-dire le nombre de probe sets.

Une autre approche du problème des tests multiples est de prendre en compte le FDR (de l'anglais False-Discovery Rate) qui est la proportion de faux positifs dans les gènes identifiés comme exprimés de manière différentielle [194]. Il existe des programmes qui calculent le FDR à partir des données de la puce à ADN.

Cependant, une correction pour tests multiples diminue le seuil de significativité à un niveau si bas que le nombre de faux négatifs augmente considérablement. Le problème reste donc fondamentalement sans solution et un grand nombre de réplicats sont nécessaires pour gagner en puissance prédictive positive et négative.

3.2.9. Interprétation des résultats

Une fois que toutes les étapes précédemment décrites ont été réalisées, on obtient une liste de gènes candidats de laquelle il faut tirer des hypothèses quant au processus biologique étudié. Comme cette tâche peut s'avérer ardue, des outils se sont développés récemment.

L'outil le plus utilisé pour interpréter des résultats venant d'expériences utilisant des puces à ADN est DAVID (Database for Annotation, Visualization and Integrated Discovery) [195, 196]. Il s'agit d'une application librement disponible sur Internet qui permet, entre autres, à partir d'une liste de gènes, de déterminer dans quelles voies particulières et chez quel organisme ces gènes sont impliqués. Pour cela, divers paramètres, comme le nombre de gènes impliqués dans une voie, doivent être choisis par l'utilisateur. Les gènes sont alors mis en évidence dans des représentations graphiques venant de bases de données telles que KEGG ou Biocarta. Ainsi, en quelques étapes, il est possible de déterminer quelles grandes voies ont été activées, réprimées ou modifiées dans une condition expérimentale donnée.

Très récemment, des outils alternatifs ont été mis au point. Notamment au sein de notre laboratoire, gViz (Helaers *et al.*, accepté dans BMC Research Notes) est un outil graphique qui, à partir des valeurs de co-expression des gènes (venant, par exemple, d'une étude utilisant des puces à ADN), permet de construire des réseaux entre ceux-ci. Plus le coefficient de corrélation liant deux gènes est élevé, plus ceux-ci vont être proches dans le réseau et plus le lien qui les unit sera fort. Cela permet de construire des réseaux de gènes qui peuvent être comparés aux cartes venant de KEGG et de Biocarta. Les différences apparaissant entre ces réseaux et les cartes des processus biologiques bien connus permettent ainsi de générer de nouvelles hypothèses. Couplé à la base de données PathEx [197], également développée dans notre laboratoire, et qui permet de trouver des jeux de données relatifs à un ou plusieurs processus biologiques choisis par l'utilisateur et de construire des jeux de données composites, gViz est un outil prometteur dans l'exploitation de l'information venant des expériences utilisant des puces à ADN.

3.3. La méta-analyse des résultats issus de puces à ADN

Les puces à ADN, et en particulier les GeneChips Affymetrix, sont donc souvent utilisées pour mesurer l'abondance des transcrits présents dans des échantillons biologiques à un moment donné. Cependant, les problèmes liés à leur analyse, et évoqués plus haut, rendent souvent les résultats ininterprétables ou même erronés. Étant donné que les données brutes peuvent désormais être stockées en format numérique, des bases de données publiques sont apparues et la ré-analyse de jeux de données est devenue pratique courante [198]. En outre, un nombre croissant d'articles décrivent des études combinant plusieurs jeux de données. Des méthodes spécifiques pour ces méta-analyses sont, à l'heure actuelle, publiées régulièrement [199-202]. Alors que certaines d'entre elles sont des méta-analyses de grande envergure,

d'autres ciblent des questions plus spécifiques, notamment dans le domaine de l'oncologie. Ces méthodes présentent l'avantage d'augmenter la puissance statistique de l'analyse et donc de fournir une cohérence et une validité biologiques plus grandes.

Par exemple, Bala Gur-Dedeoglu et ses collaborateurs [199] ont montré que, par une méthode de ré-échantillonnage dans deux jeux de données concernant le cancer du sein, il était possible de différencier tissu mammaire sain et tissu mammaire cancéreux, grâce à l'établissement d'un groupe de gènes faiblement différentiellement exprimés, qu'une méthode d'analyse classique n'aurait pas pu mettre en évidence. De la même manière, Shuangge Ma et Jian Huang [200] ont développé une approche, appelée MTGDR (pour « Meta Threshold Gradient Descent Regularization »), qui a été capable de mettre en évidence un petit nombre de gènes, dans plusieurs jeux de données, ayant un effet significatif dans le développement des cancers du pancréas et du foie. L'avantage de MTGDR est d'être robuste aux différentes conditions expérimentales des jeux de données qu'il inclut. En effet, les auteurs ont appliqué leur méthode à des jeux de données générés par différents laboratoires et utilisant des technologies différentes. Cependant, il faut rester conscient que le nombre de jeux de données que peut inclure MTGDR reste faible. En effet, les auteurs n'ont appliqué leur méthode qu'à quatre jeux de données. En 2009, Scott Ochsner et ses collaborateurs [201] mettaient à disposition un outil internet, GEMS (pour « Gene Expression MetaSignatures »), capable de réaliser la méta-analyse de plusieurs jeux de données. Ayant appliqué leur méthode à dix jeux de données traitant de la stimulation de cellules cancéreuses (de la lignée MCF-7) par une hormone (17beta-estradiol), ces auteurs ont identifié un ensemble de gènes dont 85% étaient déjà connus dans la littérature pour être impliqués dans la réponse des cellules à cette hormone. Enfin, en même temps qu'Ochsner, Andrew Sims et ses collaborateurs [202] démontraient qu'en combinant six jeux de données Affymetrix comparant des cellules de cancer du sein et des cellules de tissu mammaire sain, ils étaient capables d'augmenter la précision du pronostic des patients atteints de cette pathologie.

Ces différents exemples montrent que la méta-analyse de jeux de données provenant de puces à ADN est une méthode prometteuse pour tirer de nouvelles hypothèses quant aux mécanismes sous-jacents. Ce type de méthode assure que les gènes sélectionnés ont une haute probabilité d'être effectivement exprimés de manière différentielle. De plus, le nombre de faux positifs et de faux négatifs est substantiellement diminué. Enfin, une méta-analyse peut être réalisée à partir de jeux de données existant déjà, sans avoir besoin d'effectuer de coûteuses expériences.

Cependant, la méta-analyse n'est pas sans défaut. D'abord, la qualité des résultats d'une méta-analyse est dépendante de la qualité des analyses qu'elle prend en compte. Une bonne méthode de méta-analyse appliquée à des études de mauvaise qualité donnera des résultats médiocres. De plus, il existe un biais dans la publication des données. En effet, certaines études ne sont pas publiées parce qu'elles n'apportent pas les résultats escomptés. Or, celles-ci devraient être considérées dans le cadre de la méta-analyse. Ce biais a pour conséquence la surestimation de la significativité des études qui ont été publiées [203, 204]. Enfin, il faut rester conscient qu'une méta-analyse peut être menée dans un contexte de conflit d'intérêt. D'abord le choix des études retenues pour la méta-analyse peut être volontairement biaisé, ensuite les études peuvent elles-mêmes être biaisées à cause d'un conflit d'intérêt. Un récent article a montré que parmi 509 études reprises dans 29 méta-analyses du domaine pharmaco-médical, 132 présentaient un biais dû à des conflits d'intérêt [205]. Malgré le gain de puissance statistique que permet une méta-analyse, il convient donc de rester prudent lors du choix des études qui composeront cette méta-analyse.

3.4. Conclusion de la problématique méthodologique

Les puces à ADN sont donc une technique récente permettant, notamment, l'étude du transcriptome. Physiquement, ce sont des supports sur lesquels se trouvent, à haute densité, des sondes composées de nucléotides permettant, à l'heure actuelle, l'analyse de l'expression de gènes de génomes entiers. Grâce à la capacité de l'ARN de se lier à la séquence d'ADN qui lui est anti-complémentaire, on peut mesurer l'abondance relative des différents ARNm se trouvant à un instant donné dans un échantillon biologique dans une condition expérimentale donnée par rapport à une autre. Au fil du temps, plusieurs types de puces et de nombreuses applications se sont développés.

L'analyse des puces à ADN se fait en plusieurs étapes. Étant donné le grand nombre d'informations générées par cette technique, dite à haut débit, chacune de ces étapes nécessite des approches bioinformatiques. Et bien que ces étapes aient toutes posé des challenges, la communauté scientifique a toujours su apporter des solutions plus ou moins satisfaisantes pour obtenir, au final, des résultats biologiques validables. Ce dynamisme est nécessaire car la masse de données brutes générées par les différentes équipes est immense et publiquement disponible, grâce aux bases de données que l'on trouve sur Internet, pour la ré-analyse.

En effet, la première étape dans l'analyse de puces à ADN est la numérisation de l'image obtenue. Ces données numériques peuvent être stockées en grande quantité dans les

bases de données, donnant l'opportunité à quiconque de les ré-analyser en utilisant des choix méthodologiques différents. Effectivement, que ce soit, entre autres, dans le choix du CDF, de la méthode de prétraitement des données, ou du test statistique à effectuer, il existe de nombreuses possibilités.

Il existe donc une réelle demande pour une méthode de ré-analyse des données déjà publiées qui utilise les outils les plus récents et les plus performants afin d'optimiser les résultats biologiques qui en découlent. La méta-analyse, c'est-à-dire la combinaison de plusieurs jeux de données en une seule analyse, pourrait être la réponse à cette demande. En effet, en augmentant le nombre de jeux de données relatifs à une problématique particulière, on compense le manque de réplicats qui caractérise la plupart des expériences utilisant des puces à ADN. Ainsi on gagne en puissance statistique, qui est l'un des problèmes majeurs (si ce n'est le plus important) liés à l'analyse de puces à ADN.

4. Objectif

Dans la première partie de ce manuscrit, nous avons replacé le sujet dans son contexte. Nous avons vu que la combinaison d'outils performants et de plusieurs jeux de données issus d'expériences utilisant des puces à ADN pouvait permettre de générer de nouvelles hypothèses concernant une problématique biologique. Pour cela une méthodologie innovante doit être mise sur pied. Appliquée au processus de métastatisation des cellules cancéreuses, cette méthodologie permettrait de mettre en évidence des gènes qui n'ont pas encore été décrits comme participant à ce phénomène. Par extension, ces gènes pourraient faire partie d'une voie de signalisation qui, elle non plus, n'a pas encore été décrite comme étant impliquée dans les métastases. Sachant que toute nouvelle découverte sur le sujet est une nouvelle possibilité de combattre une maladie qui fait encore trop de décès, les enjeux de ce projet sont de première importance.

L'un des objectifs de ce projet est de sélectionner par l'analyse de puces à ADN des gènes et des voies de signalisation participant à la migration des cellules cancéreuses. Pour atteindre cet objectif, la première étape est de sélectionner des jeux de données issus d'expériences utilisant des puces à ADN relatives à l'étude des métastases et de la réponse à l'hypoxie puisque ces deux processus sont intimement liés. Il s'agit ensuite, par une méthodologie originale, d'extraire de nouvelles informations de ces puces à ADN. Ces informations doivent générer des hypothèses quant à l'implication de tel ou tel gène ou voie de signalisation dans le phénotype métastatique. Enfin dans un dernier temps, les hypothèses

générées *in silico* doivent être validées par des approches expérimentales. Ainsi, nous espérons identifier de nouvelles cibles pour la thérapie contre le cancer et particulièrement contre le développement des métastases.

Dans la seconde partie de ce manuscrit, nous allons décrire les différentes méthodes que nous avons utilisées pour obtenir les résultats qui vont être exposés ensuite. Les premières méthodes sont de type informatique. Il s'agit, notamment, des scripts et de leur description qui ont permis la méta-analyse des puces à ADN. Ensuite, viendront les protocoles de laboratoire nécessaires aux validations *in vitro*.

Dans la troisième partie, nous allons décrire les résultats des différentes expériences qui ont été menées, tant au niveau bioinformatique qu'au niveau biologie cellulaire. Les premiers résultats sont ceux qui ont permis d'élaborer la méthodologie bioinformatique pour la méta-analyse des données issues de puces à ADN. Les résultats suivants sont ceux issus de l'application de cette méta-analyse aux jeux de données étudiant le phénotype métastatique et la réponse à l'hypoxie. Et enfin, les derniers résultats qui seront présentés sont les validations que nous avons effectuées en laboratoire sur cellules cancéreuses concernant les hypothèses générées par la méthodologie bioinformatique.

Parmi ces résultats, plusieurs se trouvent sous la forme d'articles qui ont été publiés dans les revues scientifiques « BMC Cancer » et « Journal Of Proteomics And Bioinformatics ». Les résultats de la dernière partie seront également présentés sous la forme d'un article.

Enfin, nous terminerons ce manuscrit par une discussion des résultats, les perspectives que laisse cette étude et une brève conclusion reprenant les points importants que nous aurons soulignés tout au long de ce travail.

II. MATERIEL & METHODES

Dans cette partie du manuscrit, nous allons décrire en détail les différentes techniques utilisées durant ce projet. Cette partie a donc pour but de permettre à toute personne de réaliser les expériences qui ont été menées pour obtenir les résultats présentés dans la partie suivante. D'abord, nous allons décrire les jeux de données réalisés à partir de puces à ADN qui ont été utilisés. Nous expliquerons ensuite les procédures informatiques ayant permis de tester les CDFs d'AffyProbeMiner, ainsi que de réaliser la méta-analyse des puces à ADN. Enfin, nous passerons en revue les différentes techniques qui ont été utilisées pour valider *in vitro* les hypothèses bioinformatiques.

1. Sélection des jeux de données

Comme ce travail avait pour objectif de mettre en évidence l'implication de gènes dans le phénotype métastatique de cellules cancéreuses régulé par l'hypoxie, ce sont des jeux de données comparant des cellules normales ou cancéreuses à des cellules cancéreuses migratoires ou de métastases et des cellules incubées en condition hypoxique à des cellules incubées en condition normale qui ont été recherchés. De plus, comme l'expertise du laboratoire est centrée sur la technologie Affymetrix, les jeux de données qui ont été retenus utilisent les puces à ADN de cette marque.

Tous les jeux de données ont été téléchargés à partir des bases de données GEO et ArrayExpress au format .CEL. Pour les jeux de données qui n'étaient pas publiquement disponibles, les auteurs ont été contactés. Les jeux de données contenant plus d'un modèle de puce à ADN et/ou plus de deux conditions expérimentales ont été divisés pour répondre à ce critère. Au total, ce sont 22 jeux de données qui ont été utilisés (Tableau 1).

Tableau 1. *Le numéro d'accès de GEO et de ArrayExpress (AE) avec le modèle de GeneChip correspondant ainsi que les conditions expérimentales.*

Numéros d'accès	Modèles de GeneChip	Bases de données	Disponibilité	Conditions expérimentales
E-GEOD-1323	HG-U133A	AE	Disponible	3 échantillons de tumeur primaire colorectale VS. 3 échantillons de métastases au niveau des ganglions lymphatiques
E-GEOD-2280	HG-U133A	AE	Disponible	8 échantillons de carcinome de la cavité orale VS. 19 échantillons de métastases au niveau des ganglions lymphatiques
E-MEXP-44	HG-U95Av2 HG-UgeneFL	AE	Disponible	15 échantillons de carcinome au niveau de la tête et du cou VS. 3 échantillons de métastases au niveau des ganglions lymphatiques 12 échantillons de carcinome au niveau de la tête et du cou VS. 11 échantillons de métastases au niveau des ganglions lymphatiques

Numéros d'accès	Modèles de GeneChip	Bases de données	Disponibilité	Conditions expérimentales
GSE1056	HG-U95Av2	GEO	Non disponible	2 échantillons de carcinome hépatique en hypoxie pendant 2 h VS. 2 échantillons de carcinome hépatique en normoxie 2 échantillons de carcinome hépatique en hypoxie pendant 24 h VS. 2 échantillons de carcinome hépatique en normoxie
GSE2280	HG-U133A	GEO	Disponible	22 échantillons de carcinome de la cavité orale VS. 5 échantillons de métastases au niveau des ganglions lymphatiques
GSE2603	HG-U133A	GEO	Disponible	100 échantillons de tumeur primaire de cancer du sein VS. 21 échantillons de métastases au niveau des poumons
GSE3325	HG-U133Plus2.0	GEO	Disponible	7 échantillons de tumeur primaire de cancer de la prostate VS. 6 métastases
GSE4086	HG-U133Plus2.0	GEO	Disponible	2 échantillons de lymphome de Burkitt en hypoxie VS. 2 échantillons de lymphome de Burkitt en normoxie
GSE468	HC-G110	GEO	Disponible	13 échantillons de médulloblastome primaire VS. 10 échantillons de métastases de médulloblastome
GSE4840	HG-U133A HG-U133B	GEO	Non disponible	3 échantillons de culture de mélanocytes VS. 12 échantillons de culture de métastases de mélanomes 3 échantillons de culture de mélanocytes VS. 12 échantillons de culture de métastases de mélanomes
GSE4843	HG-U133Plus2.0	GEO	Non disponible	45 échantillons de culture de métastases de mélanomes
GSE6369	HG-U133Plus2.0	GEO	Disponible	1 échantillon de tumeur primaire de cancer de la prostate VS. 1 échantillon de métastases de cancer de la prostate
GSE6919	HG-U95Av2 HG-U95B HG-U95C	GEO	Disponible	65 échantillons de tumeur primaire de cancer de la prostate VS. 25 échantillons de métastases de cancer de la prostate 66 échantillons de tumeur primaire de cancer de la prostate VS. 25 échantillons de métastases de cancer de la prostate 65 échantillons de tumeur primaire de cancer de la prostate VS. 25 échantillons de métastases de cancer de la prostate
GSE7929	HG-U133A	GEO	Disponible	11 échantillons de mélanome peu métastatique VS. 21 échantillons de mélanome très métastatique
GSE7930	HG-U133A	GEO	Disponible	3 échantillons de tumeur peu métastatique de cancer de la prostate VS. 3 échantillons de tumeur très métastatique de cancer de la prostate
GSE7956	HG-U133A	GEO	Disponible	10 échantillons de mélanome peu métastatique VS. 29 échantillons de mélanome très métastatique
GSE8401	HG-U133A	GEO	Disponible	31 échantillons de mélanome primaire VS. 52 échantillons de métastases de mélanome

2. Ressources informatiques

Toutes les approches bioinformatiques décrites ci-après (comparaison des CDFs, analyses individuelles, intersections, intersections d'unions, méta-analyses et profiles d'expression) ont été réalisées à l'aide du programme statistique R, dans ses versions 2.4.0, 2.6.0 et 2.10.1, et de packages venant de Bioconductor et d'AffyProbeMiner. Les scripts ont été exécutés sur un ordinateur 64-bit avec 4 gb de DDR (biprocesseur dual-core Xeon 5160 3.0Ghz, 8 x 500gb RAID).

3. Comparaison des CDFs standard et alternatif

Afin de montrer l'effet du changement de CDF sur les résultats d'une analyse de puces à ADN, un jeu de données (E-GEOD-1323) a été analysé deux fois. La première fois en utilisant le CDF standard d'Affymetrix, le prétraitement de RMA et le traitement statistique du test t de Student. La seconde en utilisant le CDF alternatif d'AffyProbeMiner, le prétraitement de RMA et le traitement statistique du test t de Student. Ainsi, seul le CDF différait entre les deux analyses (Scripts 1 et 2). Celles-ci sont présentées plus en détail dans la partie « résultats » de ce manuscrit.

Script 1. *Analyse du jeu de données E-GEOD-1323 avec un CDF standard d'Affymetrix, le prétraitement de RMA et le traitement statistique du test t de Student.*

```
setwd("/…/…/E-GEOD-1323")
library(Biobase)
library(affy)
library(rma)
library(pegase)
a<-ReadAffy()
b<-rma(a)
d<-exprs(b)
e<-d[,(1:3)]
f<-d[,(4:6)]
d2<-d*log10(2)
e2<-d2[,(1:3)]
f2<-d2[,(4:6)]
g<-apply(e2,1,mean)
h<-apply(f2,1,mean)
j<-g-h
k<-pegase(A=e,B=f,steps=c("prepare","run"),methods=c("student"))
l<--log10(k$pvals)
plot(j,l)
```

Script 2. *Analyse du jeu de données E-GEOD-1323 avec un CDF alternatif d'AffyProbeMiner, le prétraitement de RMA et le traitement statistique du test t de Student.*

```
setwd("/…/…/E-GEOD-1323")
library(hgu133acdf)
library(Biobase)
library(affy)
library(rma)
library(pegase)
library(hgu133atranscriptccds)
data(hgu133atranscriptccds)
a<-ReadAffy()
hgu133atranscriptccdsdim<-hgu133adim
a@cdfName<-"hgu133atranscriptccds"
b<-rma(a)
d<-exprs(b)
e<-d[,(1:3)]
f<-d[,(4:6)]
d2<-d*log10(2)
e2<-d2[,(1:3)]
f2<-d2[,(4:6)]
```

```
g<-apply(e2,1,mean)
h<-apply(f2,1,mean)
j<-g-h
k<-pegase(A=e,B=f,steps=c("prepare","run"),methods=c("student"))
l<--log10(k$pvals)
plot(j,l)
```

4. Analyses individuelles

Avant de réaliser la méta-analyse, les jeux de données venant d'expériences utilisant des puces à ADN ont dû être analysés individuellement (Script 3). Pour cela, des CDFs d'AffyProbeMiner ont été utilisés. Un CDF est nécessaire par modèle de puce et trois packages sont nécessaires par CDF. Les packages « CDF distribution » ont été utilisés dans leur version 1.8.0. Les packages « PROBE distribution » ont été utilisés dans leur version 1.0.0. Les packages « Annotation distribution » ont été utilisés dans leur version 1.1.0. Les CDFs utilisés étaient « transcript-consistent », ainsi les sondes d'un même probe set correspondaient toutes au même groupe de transcrits d'un gène. Les CDFs utilisés ont été conçus grâce à l'information des bases de données RefSeq et GeneBank. La taille minimale d'un probe set a été fixée à 5 sondes comme l'ont recommandé Liu et ses collègues. Le prétraitement a été réalisé avec GCRMA. Le traitement statistique a été réalisé avec le Window t test avec une correction de Welch grâce au package Pegase. Le jeu de données GSE4843 n'a pas été analysé car il ne présentait qu'une seule condition expérimentale. Le jeu de données GSE6369 n'a pas été analysé car il ne présentait qu'un seul réplicat par condition expérimentale.

Script 3. *Le GeneChip HG-U133A a été utilisé pour cet exemple de script. Certains objets et certaines valeurs, symbolisés ici par X ou Y, doivent être remplacés en fonction du jeu de données analysé.*

```
setwd("/…/…/DataSet")
library(hgu133acdf)
library(gcrma)
library(pegase)
library(hgu133atranscriptccds)
data(hgu133atranscriptccds)
hgu133atranscriptccdsdim<-hgu133adim
a<-justGCRMA(cdfname="hgu133atranscriptccds")
b<-exprs(a)
d<-b[,(1:X)]
e<-b[,(X+1:Y)]
f<-pegase(A=d,B=e,steps=c("prepare","run"),methods=c("win.welch"))
```

Les informations résultant des analyses individuelles ont ensuite été placées dans des data frames. Chaque jeu de données a ainsi permis de générer un data frame. Un data frame est un tableau dans lequel des données peuvent être liées. Ainsi, chaque gène (désigné par son identifiant de la base de données Entrez) représenté sur la puce à ADN a été lié au probe set et à la p value lui correspondant. Comme certains probe sets correspondent à plusieurs gènes, ces probe sets particuliers sont répétés plusieurs fois dans le data frame avec à chaque fois un gène différent lui étant lié, ainsi que la p value correspondante (Script 4).

Script 4. *Le modèle de GeneChip HG-U133A a été utilisé pour cet exemple de script. Certains objets et certaines valeurs comme la longueur de certains vecteurs doivent être remplacés en fonction du modèle de GeneChip analysé.*

```
load("/…/…/DataSet")
library(hgu133atranscriptccds)
data(hgu133atranscriptccds)
library(hgu133atranscriptccdscdf)
a<-mget(ls(env=hgu133atranscriptccdscdf),hgu133atranscriptccdsENTREZGENEID)
b<-c()
for(i in 1:15278){
b[i]<-a[[i]][1]}
d<-c()
for(i in 1:15278){
d[i]<-a[[i]][2]}
e<-which(d>0)
d<-d[!is.na(d)]
f<-c()
for(i in 1:15278){
f[i]<-a[[i]][3]}
g<-which(f>0)
f<-f[!is.na(f)]
h<-c()
for(i in 1:15278){
h[i]<-a[[i]][4]}
j<-which(h>0)
h<-h[!is.na(h)]
k<-c()
for(i in 1:15278){
k[i]<-a[[i]][5]}
l<-which(k>0)
k<-k[!is.na(k)]
m<-c()
for(i in 1:15278){
m[i]<-a[[i]][6]}
n<-which(m>0)
m<-m[!is.na(m)]
o<-c()
for(i in 1:15278){
o[i]<-a[[i]][7]}
p<-which(o>0)
o<-o[!is.na(o)]
q<-c()
for(i in 1:15278){
q[i]<-a[[i]][8]}
r<-which(q>0)
q<-q[!is.na(q)]
```

```
geneID<-c(b,d,f,h,k,m,o,q)
a<-names(PV)
b<-a[e]
d<-a[g]
f<-a[j]
h<-a[l]
k<-a[n]
m<-a[p]
o<-a[r]
probeset<-c(a,b,d,f,h,k,m,o)
a<-PV
b<-a[e]
d<-a[g]
f<-a[j]
h<-a[l]
k<-a[n]
m<-a[p]
o<-a[r]
PValue<-c(a,b,d,f,h,k,m,o)
DF<-data.frame(probeset,geneID,PValue)
```

5. Intersections

Les analyses individuelles ont fourni une liste de gènes pour chaque jeu de données. Pour chacune de ces listes, les gènes ont été classés selon l'ordre croissant de leur p value de leur expression différentielle. Ainsi, chaque gène occupait un rang particulier dans la liste ; les gènes étant le plus significativement sur- ou sous-exprimés se trouvaient en haut de la liste.

33 groupes de jeux de données ont été conçus (Tableau 2). Un rang seuil a été calculé pour chaque groupe grâce à la formule (1) :

$$r = [1 - (1 - P)^{1/n}]^{1/k} \times N \qquad (1)$$

où r = le rang seuil, P = la probabilité fixée, n = le nombre de gène suspectés d'être impliqués dans les processus de métastase et/ou de réponse à l'hypoxie, k = le nombre de jeux de données dans le groupe et N = le nombre de probe sets présents sur la puce à ADN (le plus grand quand plusieurs modèles de GeneChip Affymetrix sont impliqués dans le groupe).

Tableau 2. *Les 33 groupes de jeux de données pour les intersections.*

	Conditions expérimentales	**Modèles de GeneChip**	**Jeux de données**
Groupe 1	Tumeur primaire VS. métastase	HG-U133A	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE7929, GSE7956, GSE8401
Groupe 2	Tumeur primaire VS. métastase, tissu normal VS. métastase	HG-U133A	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE4840 (HG-U133A), GSE7929, GSE7956, GSE8401
Groupe 3	Tumeur primaire VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE7929, GSE7930, GSE7956, GSE8401

	Conditions expérimentales	Modèles de GeneChip	Jeux de données
Groupe 4	Tumeur primaire VS. métastase, tissu normal VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE4840 (HG-U133A), GSE7929, GSE7930, GSE7956, GSE8401
Groupe 5	Tumeur primaire VS. métastase	HG-U95Av2	E-MEXP-44 (HG-U95Av2), GSE6919 (HG-U95Av2)
Groupe 6	Tumeur primaire VS. métastase	HG-U133A, HG-U95Av2	E-GEOD-1323, E-GEOD-2280, E-MEXP-44 (HG-U95Av2), GSE2280, GSE2603, GSE6919 (HG-U95Av2), GSE7929, GSE7956, GSE8401
Groupe 7	Tumeur primaire VS. métastase, tissu normal VS. métastase	HG-U133A, HG-U95Av2	E-GEOD-1323, E-GEOD-2280, E-MEXP-44 (HG-U95Av2), GSE2280, GSE2603, GSE4840 (HG-U133A), GSE6919 (HG-U95Av2), GSE7929, GSE7956, GSE8401
Groupe 8	Tumeur primaire VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A, HG-U95Av2	E-GEOD-1323, E-GEOD-2280, E-MEXP-44 (HG-U95Av2), GSE2280, GSE2603, GSE6919 (HG-U95Av2), GSE7929, GSE7930, GSE7956, GSE8401
Groupe 9	Tumeur primaire VS. métastase, tissu normal VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A, HG-U95Av2	E-GEOD-1323, E-GEOD-2280, E-MEXP-44 (HG-U95Av2), GSE2280, GSE2603, GSE4840 (HG-U133A), GSE6919 (HG-U95Av2), GSE7929, GSE7930, GSE7956, GSE8401
Groupe 10	Tumeur primaire VS. métastase	HG-U133A, HG-U133Plus2.0	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE3325, GSE7929, GSE7956, GSE8401
Groupe 11	Tumeur primaire VS. métastase, tissu normal VS. métastase	HG-U133A, HG-U133Plus2.0	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE3325, GSE4840 (HG-U133A), GSE7929, GSE7956, GSE8401
Groupe 12	Tumeur primaire VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A, HG-U133Plus2.0	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE3325, GSE7929, GSE7930, GSE7956, GSE8401
Groupe 13	Tumeur primaire VS. métastase, tissu normal VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A, HG-U133Plus2.0	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE3325, GSE4840 (HG-U133A), GSE7929, GSE7930, GSE7956, GSE8401
Groupe 14	Tumeur primaire VS. métastase	HG-U133Plus2.0, HG-U95Av2	E-MEXP-44 (HG-U95Av2), GSE3325, GSE6919 (HG-U95Av2)
Groupe 15	Tumeur primaire VS. métastase	HG-U133A, HG-U133Plus2.0, HG-U95Av2	E-GEOD-1323, E-GEOD-2280, E-MEXP-44 (HG-U95Av2), GSE2280, GSE2603, GSE3325, GSE6919 (HG-U95Av2), GSE7929, GSE7956, GSE8401
Groupe 16	Tumeur primaire VS. métastase, tissu normal VS. métastase	HG-U133A, HG-U133Plus2.0, HG-U95Av2	E-GEOD-1323, E-GEOD-2280, E-MEXP-44 (HG-U95Av2), GSE2280, GSE2603, GSE3325, GSE4840 (HG-U133A), GSE6919 (HG-U95Av2), GSE7929, GSE7956, GSE8401
Groupe 17	Tumeur primaire VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A, HG-U133Plus2.0, HG-U95Av2	E-GEOD-1323, E-GEOD-2280, E-MEXP-44 (HG-U95Av2), GSE2280, GSE2603, GSE3325, GSE6919 (HG-U95Av2), GSE7929, GSE7930, GSE7956, GSE8401
Groupe 18	Tumeur primaire VS. métastase, tissu normal VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A, HG-U133Plus2.0, HG-U95Av2	E-GEOD-1323, E-GEOD-2280, E-MEXP-44 (HG-U95Av2), GSE2280, GSE2603, GSE3325, GSE4840 (HG-U133A), GSE6919 (HG-U95Av2), GSE7929, GSE7930, GSE7956, GSE8401
Groupe 19	Tumeur primaire VS. métastase	HG-U133B, HG-U95B	GSE4840 (HG-U133B), GSE6919 (HG-U95B)
Groupe 20	Normoxie VS. hypoxie	HG-U95Av2	GSE1056
Groupe 21	Tumeur primaire VS. métastase	HG-U133A, HG-U133Plus2.0, HG-U95Av2, HG-U95B, HG-U95C, HU-geneFL, HC-G110	E-GEOD-1323, E-GEOD-2280, E-MEXP-44, GSE2280, GSE2603, GSE3325, GSE468, GSE6919, GSE7929, GSE7956, GSE8401

	Conditions expérimentales	Modèles de GeneChip	Jeux de données
Groupe 22	Tumeur primaire VS. métastase, tissu normal VS. métastase	All GeneChip models	E-GEOD-1323, E-GEOD-2280, E-MEXP-44, GSE2280, GSE2603, GSE3325, GSE468, GSE4840, GSE6919, GSE7929, GSE7956, GSE8401
Groupe 23	Tumeur primaire VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A, HG-U133Plus2.0, HG-U95Av2, HG-U95B, HG-U95C, HU-geneFL, HC-G110	E-GEOD-1323, E-GEOD-2280, E-MEXP-44, GSE2280, GSE2603, GSE3325, GSE468, GSE6919, GSE7929, GSE7930, GSE7956, GSE8401
Groupe 24	Tumeur primaire VS. métastase, tissu normal VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	All GeneChip models	E-GEOD-1323, E-GEOD-2280, E-MEXP-44, GSE2280, GSE2603, GSE3325, GSE468, GSE4840, GSE6919, GSE7929, GSE7930, GSE7956, GSE8401
Groupe 25	Normoxie VS. hypoxie	HG-U133Plus2.0, HG-U95Av2	GSE1056, GSE4086
Groupe 26	Carcinome de la cavité orale VS. métastase au niveau des ganglions lymphatiques	HG-U133A	E-GEOD-2280, GSE2280
Groupe 27	Carcinome au niveau de la tête et du cou VS. métastase au niveau des ganglions lymphatiques	HG-U95Av2, HU-geneFL	E-MEXP-44
Groupe 28	Tumeur primaire de la prostate VS. métastase	HG-U133Plus2.0, HG-U95B, HG-U95C	GSE3325, GSE6919 (HG-U95B, HG-U95C)
Groupe 29	Tumeur primaire de la prostate VS. métastase, tumeur peu métastatique de la prostate VS. tumeur très métastatique de la prostate	HG-U133A, HG-U133Plus2.0, HG-U95B, HG-U95C	GSE3325, GSE6919 (HG-U95B, HG-U95C), GSE7930
Groupe 30	Mélanome primaire VS. métastase de mélanome, mélanome peu métastatique VS. mélanome très métastatique	HG-U133A	GSE7929, GSE7956, GSE8401
Groupe 31	Culture de mélanocytes normaux VS. culture de cellules de metastase de mélanome	HG-U133A, HG-U133B	GSE4840
Groupe 32	Mélanome primaire VS. métastase de mélanome, mélanome peu métastatique VS. mélanome très métastatique, culture de mélanocytes normaux VS. culture de cellules de metastase de mélanome	HG-U133A, HG-U133B	GSE4840, GSE7929, GSE7956, GSE8401
Groupe 33	Toutes les conditions	Tous les modèles de GeneChip	Tous les jeux de données

Les gènes communs à tous les jeux de données du groupe et ayant un rang inférieur au rang seuil ont été sélectionnés (Script 5).

Script 5. *Certains objets et certaines valeurs, symbolisés ici par X, doivent être remplacés en fonction des jeux de données impliqués dans l'intersection. DF signifie data frame.*

```
load("/…/DF1")
load("/…/DF2")
load("/…/DF3")
a<-DF1[order(DF1$PValue),]
b<-DF2[order(DF2$PValue),]
d<-DF3[order(DF3$PValue),]
e<-a$geneID[1:X]
f<-b$geneID[1:X]
g<-d$geneID[1:X]
a<-intersect(e,f)
b<-intersect(g,a)
```

6. Intersections d'unions

Comme pour les intersections, les gènes ont été classés selon l'ordre croissant de leur p value. Les 17 jeux de données relatifs aux métastases ont été assemblés en 30 groupes différents. Les 3 jeux de données relatifs à la réponse à l'hypoxie ont été assemblés en un

groupe (Tableau 3). Chaque groupe relatif aux métastases a été pris en considération avec le groupe relatif à l'hypoxie. Pour chaque couple de groupes, les 50 gènes les plus significatifs communs à au moins un jeu de données portant sur les métastases et à au moins un jeu de données portant sur l'hypoxie ont été sélectionnés (Script 6).

Tableau 3. *Les 30 groupes de jeux de données relatifs aux métastases et le groupe de jeux de données relatif à l'hypoxie pour les intersections d'unions.*

	Conditions expérimentales	**Modèles de GeneChip**	**Jeux de données**
Groupe métastase 1	Tumeur primaire VS. métastase	HG-U133A	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE7929, GSE7956, GSE8401
Groupe métastase 2	Tumeur primaire VS. métastase, tissu normal VS. métastase	HG-U133A	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE4840 (HG-U133A), GSE7929, GSE7956, GSE8401
Groupe métastase 3	Tumeur primaire VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE7929, GSE7930, GSE7956, GSE8401
Groupe métastase 4	Tumeur primaire VS. métastase, tissu normal VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE4840 (HG-U133A), GSE7929, GSE7930, GSE7956, GSE8401
Groupe métastase 5	Tumeur primaire VS. métastase	HG-U95Av2	E-MEXP-44 (HG-U95Av2), GSE6919 (HG-U95Av2)
Groupe métastase 6	Tumeur primaire VS. métastase	HG-U133A, HG-U95Av2	E-GEOD-1323, E-GEOD-2280, E-MEXP-44 (HG-U95Av2), GSE2280, GSE2603, GSE6919 (HG-U95Av2), GSE7929, GSE7956, GSE8401
Groupe métastase 7	Tumeur primaire VS. métastase, tissu normal VS. métastase	HG-U133A, HG-U95Av2	E-GEOD-1323, E-GEOD-2280, E-MEXP-44 (HG-U95Av2), GSE2280, GSE2603, GSE4840 (HG-U133A), GSE6919 (HG-U95Av2), GSE7929, GSE7956, GSE8401
Groupe métastase 8	Tumeur primaire VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A, HG-U95Av2	E-GEOD-1323, E-GEOD-2280, E-MEXP-44 (HG-U95Av2), GSE2280, GSE2603, GSE6919 (HG-U95Av2), GSE7929, GSE7930, GSE7956, GSE8401
Groupe métastase 9	Tumeur primaire VS. métastase, tissu normal VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A, HG-U95Av2	E-GEOD-1323, E-GEOD-2280, E-MEXP-44 (HG-U95Av2), GSE2280, GSE2603, GSE4840 (HG-U133A), GSE6919 (HG-U95Av2), GSE7929, GSE7930, GSE7956, GSE8401
Groupe métastase 10	Tumeur primaire VS. métastase	HG-U133A, HG-U133Plus2.0	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE3325, GSE7929, GSE7956, GSE8401
Groupe métastase 11	Tumeur primaire VS. métastase, tissu normal VS. métastase	HG-U133A, HG-U133Plus2.0	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE3325, GSE4840 (HG-U133A), GSE7929, GSE7956, GSE8401
Groupe métastase 12	Tumeur primaire VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A, HG-U133Plus2.0	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE3325, GSE7929, GSE7930, GSE7956, GSE8401
Groupe métastase 13	Tumeur primaire VS. métastase, tissu normal VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A, HG-U133Plus2.0	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE3325, GSE4840 (HG-U133A), GSE7929, GSE7930, GSE7956, GSE8401
Groupe métastase 14	Tumeur primaire VS. métastase	HG-U133Plus2.0, HG-U95Av2	E-MEXP-44 (HG-U95Av2), GSE3325, GSE6919 (HG-U95Av2)

	Conditions expérimentales	Modèles de GeneChip	Jeux de données
Groupe métastase 15	Tumeur primaire VS. métastase	HG-U133A, HG-U133Plus2.0, HG-U95Av2	E-GEOD-1323, E-GEOD-2280, E-MEXP-44 (HG-U95Av2), GSE2280, GSE2603, GSE3325, GSE6919 (HG-U95Av2), GSE7929, GSE7956, GSE8401
Groupe métastase 16	Tumeur primaire VS. métastase, tissu normal VS. métastase	HG-U133A, HG-U133Plus2.0, HG-U95Av2	E-GEOD-1323, E-GEOD-2280, E-MEXP-44 (HG-U95Av2), GSE2280, GSE2603, GSE3325, GSE4840 (HG-U133A), GSE6919 (HG-U95Av2), GSE7929, GSE7956, GSE8401
Groupe métastase 17	Tumeur primaire VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A, HG-U133Plus2.0, HG-U95Av2	E-GEOD-1323, E-GEOD-2280, E-MEXP-44 (HG-U95Av2), GSE2280, GSE2603, GSE3325, GSE6919 (HG-U95Av2), GSE7929, GSE7930, GSE7956, GSE8401
Groupe métastase group 18	Tumeur primaire VS. métastase, tissu normal VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A, HG-U133Plus2.0, HG-U95Av2	E-GEOD-1323, E-GEOD-2280, E-MEXP-44 (HG-U95Av2), GSE2280, GSE2603, GSE3325, GSE4840 (HG-U133A), GSE6919 (HG-U95Av2), GSE7929, GSE7930, GSE7956, GSE8401
Groupe métastase 19	Tumeur primaire VS. métastase, tissu normal VS. métastase	HG-U133B, HG-U95B	GSE4840 (HG-U133B), GSE6919 (HG-U95B)
Groupe métastase 20	Tumeur primaire VS. métastase	HG-U133A, HG-U133Plus2.0, HG-U95Av2, HG-U95B, HG-U95C, HU-geneFL, HC-G110	E-GEOD-1323, E-GEOD-2280, E-MEXP-44, GSE2280, GSE2603, GSE3325, GSE468, GSE6919, GSE7929, GSE7956, GSE8401
Groupe métastase 21	Tumeur primaire VS. métastase, tissu normal VS. métastase	All GeneChip models	E-GEOD-1323, E-GEOD-2280, E-MEXP-44, GSE2280, GSE2603, GSE3325, GSE468, GSE4840, GSE6919, GSE7929, GSE7956, GSE8401
Groupe métastase 22	Tumeur primaire VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	HG-U133A, HG-U133Plus2.0, HG-U95Av2, HG-U95B, HG-U95C, HU-geneFL, HC-G110	E-GEOD-1323, E-GEOD-2280, E-MEXP-44, GSE2280, GSE2603, GSE3325, GSE468, GSE6919, GSE7929, GSE7930, GSE7956, GSE8401
Groupe métastase 23	Tumeur primaire VS. métastase, tissu normal VS. métastase, tumeur peu métastatique VS. tumeur très métastatique	All GeneChip models	E-GEOD-1323, E-GEOD-2280, E-MEXP-44, GSE2280, GSE2603, GSE3325, GSE468, GSE4840, GSE6919, GSE7929, GSE7930, GSE7956, GSE8401
Groupe métastase 24	Carcinome de la cavité orale VS. métastase au niveau des ganglions lymphatiques	HG-U133A	E-GEOD-2280, GSE2280
Groupe métastase 25	Carcinome au niveau de la tête et du cou VS. métastase au niveau des ganglions lymphatiques	HG-U95Av2, HU-geneFL	E-MEXP-44
Groupe métastase 26	Tumeur primaire de la prostate VS. métastase	HG-U133Plus2.0, HG-U95B, HG-U95C	GSE3325, GSE6919 (HG-U95B, HG-U95C)
Groupe métastase 27	Tumeur primaire de la prostate VS. métastase, tumeur peu métastatique de la prostate VS. tumeur très métastatique de la prostate	HG-U133A, HG-U133Plus2.0, HG-U95B, HG-U95C	GSE3325, GSE6919 (HG-U95B, HG-U95C), GSE7930
Groupe métastase 28	Mélanome primaire VS. métastase de mélanome, mélanome peu métastatique VS. mélanome très métastatique	HG-U133A	GSE7929, GSE7956, GSE8401
Groupe métastase 29	Culture de mélanocytes normaux VS. culture de cellules de métastase de mélanome	HG-U133A, HG-U133B	GSE4840
Groupe métastase 30	Mélanome primaire VS. métastase de mélanome, mélanome peu métastatique VS. mélanome très métastatique, culture de mélanocytes normaux VS. culture de cellules de métastase de mélanome	HG-U133A, HG-U133B	GSE4840, GSE7929, GSE7956, GSE8401
Groupe hypoxie 1	Normoxie VS. hypoxie	HG-U95Av2, HG-U133Plus2.0	GSE1056, GSE4086

Script 6. *Certains objets et certaines valeurs, symbolisés ici par X, doivent être remplacés en fonction des jeux de données impliqués dans l'intersection d'union. DF signifie data frame.*

```
load("/…/DF1")
load("/…/DF2")
load("/…/DF3")
load("/…/DF4")
load("/…/DF5")
load("/…/DF6")
Z1<-DF1[order(DF1$PValue),]
Z2<-DF2[order(DF2$PValue),]
Z3<-DF3[order(DF3$PValue),]
Z4<-DF4[order(DF4$PValue),]
Z5<-DF5[order(DF5$PValue),]
Z6<-DF6[order(DF6$PValue),]
a<-Z1$geneID[1:X]
a<-as.character(a)
b<-Z2$geneID[1:X]
b<-as.character(b)
d<-Z3$geneID[1:X]
d<-as.character(d)
e<-c(a,b,d)
A<-Z4$geneID[1:X]
A<-as.character(A)
B<-Z5$geneID[1:X]
B<-as.character(B)
D<-Z6$geneID[1:X]
D<-as.character(D)
E<-c(A,B,D)
union<-intersect(e,E)
```

7. Méta-analyses

Les 22 jeux de données ont été assemblés en 14 jeux composites (Tableau 4). Les CDFs d'AffyProbeMiner ont été utilisés. Le prétraitement a été réalisé avec GCRMA et le traitement statistique avec le Window t test avec une correction de Welch grâce au package Pegase. Pour chaque jeu composite, les 50 gènes les plus significatifs ont été sélectionnés (Script 7).

Tableau 4. *Les 14 jeux composites.*

	Conditions expérimentales	**Modèles de GeneChip**	**Jeux de données**
Meta-jeu de données 1	Tumeur primaire, tissu normal, tumeur peu métastatique VS. métastase, tumeur très métastatique	HG-U133A	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE4840 (HG-U133A), GSE7929, GSE7930, GSE7956, GSE8401
Meta- jeu de données 2	Tumeur primaire, tumeur peu métastatique VS. métastase, tumeur très métastatique	HG-U133A	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE7929, GSE7930, GSE7956, GSE8401
Meta- jeu de données 3	Tumeur primaire, tissu normal VS. métastase	HG-U133A	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE4840 (HG-U133A), GSE7929, GSE7956, GSE8401

	Conditions expérimentales	Modèles de GeneChip	Jeux de données
Meta- jeu de données 4	Tumeur primaire VS. métastase	HG-U133A	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE4840 (HG-U133A), GSE7929, GSE7956, GSE8401
Meta- jeu de données 5	Tumeur primaire VS. métastase	HG-U133A	E-GEOD-1323, E-GEOD-2280, GSE2280, GSE2603, GSE7929, GSE7956, GSE8401
Meta- jeu de données 6	Carcinome de la cavité orale VS. métastase au niveau des ganglions lymphatiques	HG-U133A	E-GEOD-2280, GSE2280
Meta- jeu de données 7	Culture de mélanocytes normaux, mélanome peu métastatique, mélanome primaire VS. culture de cellules de métastase de mélanome, mélanome très métastatique, métastase de mélanome	HG-U133A	GSE4840 (HG-U133A), GSE7929, GSE7956, GSE8401
Meta- jeu de données 8	mélanome peu métastatique, mélanome primaire VS. culture de cellules de métastase de mélanome, mélanome très métastatique, métastase de mélanome	HG-U133A	GSE4840 (HG-U133A), GSE7929, GSE7956, GSE8401
Meta- jeu de données 9	mélanome peu métastatique, mélanome primaire VS. mélanome très métastatique, métastase de mélanome	HG-U133A	GSE7929, GSE7956, GSE8401
Meta- jeu de données 10	Tumeur primaire VS. métastase	HG-U95Av2	E-MEXP-44 (HG-U95Av2), GSE6919 (HG-U95Av2)
Meta- jeu de données 11	Normoxie VS. hypoxie	HG-U95Av2	GSE1056
Meta- jeu de données 12	Tumeur primaire, normoxie VS. métastase, hypoxie	HG-U133Plus2.0	GSE3325, GSE4086, GSE4843, GSE6369
Meta- jeu de données 13	Tumeur primaire VS. métastase	HG-U133Plus2.0	GSE3325, GSE4843, GSE6369
Meta- jeu de données 14	Tumeur primaire de la prostate VS. métastase	HG-U133Plus2.0	GSE3325, GSE6369

Script 7. *Le GeneChip HG-U133A a été utilisé pour cet exemple de script. Certains objets et certaines valeurs, symbolisés ici par X ou Y, doivent être remplacés en fonction du jeu de données analysé.*

```
setwd("/…/…/MetaDataSet")
library(hgu133acdf)
library(gcrma)
library(pegase)
library(hgu133atranscriptccds)
data(hgu133atranscriptccds)
hgu133atranscriptccdsdim<-hgu133adim
a<-justGCRMA(cdfname="hgu133atranscriptccds")
b<-exprs(a)
d<-b[,(1:X)]
e<-b[,(X+1:Y)]
f<-pegase(A=d,B=e,steps=c("prepare","run"),methods=c("win.welch"))
g<-sort(f$pvals)
h<-g[(1:50)]
```

8. Réseaux de gènes

Les gènes sélectionnés par les intersections, les intersections d'unions et les méta-analyses ont été soumis à la version 6.7 de DAVID (Database for Annotation, Visualization and Integrated Discovery). Les paramètres du « Functional Annotation Tool » ont été fixés de manière à rapatrier uniquement des voies de signalisation venant des bases de données KEGG et Biocarta car ce sont les seules à en proposer des versions graphiques. Les paramètres du « Functional Annotation Clustering » ont été fixés au niveau le plus bas afin de rapatrier le plus grand nombre possible de voies.

9. Profils d'expression

Les jeux de données GSE5823 et GSE9350 ont été téléchargés à partir de la base de données GEO. Le jeu de données GSE5823 compare 3 échantillons de culture de cellules MCF-7 et 2 échantillons de culture de cellules MDA-MB-231. Le jeu de données GSE9350 compare 3 échantillons de culture de cellules L3.6pl en condition hypoxique et 3 échantillons de culture de cellules L3.6pl en condition normale. Les deux jeux de données ont été analysés individuellement avec un CDF d'AffyProbeMiner et prétraités avec GCRMA avec les paramètres par défaut. Les valeurs d'expression ont ensuite été placées sur l'axe des ordonnées d'un graphique alors que l'axe des abscisses représentait le type cellulaire ou la condition expérimentale.

10. Cultures cellulaires

Des cellules de types MDA-MB-231 et MCF-7 ont été cultivées séparément dans du milieu de culture Roswell Park Memorial Institute (RPMI) 1640 (Gibco, Paisley, Royaume-Uni) contenant 10% de sérum de veau fœtal. Ces cultures cellulaires ont été maintenues à 37°C dans une atmosphère contenant 5% de CO_2.

11. Extraction d'ARN total

Des cellules de types MDA-MB-231 et MCF-7 ont été cultivées séparément dans des boîtes T75. Le milieu de culture a été enlevé et l'extraction d'ARN total a été réalisée à l'aide du kit RNAgents (RNAgents, Total RNA Isolation System, Promega, Madison) selon les instructions du fabricant. L'ARN total a ensuite été placé à -70°C.

12. Rétro-transcription

Pour chaque type cellulaire (MDA-MB-231 et MCF-7), 1 µg d'ARN total a été dilué dans 12 µl d'eau. 1 µl d'Anchored-oligo (dT) 18 Primer (50 pmol/µL) (Roche) a ensuite été ajouté. Ce mélange a été incubé pendant 10 minutes à 65°C. 7 µl du mélange de réaction [4 µl de Transcriptor Reverse Transcriptase Reaction Buffer 5X (Roche) ; 0,5 µl de Protector RNase Inhibitor (40 U/µl) (Roche) ; 2 µl de Deoxynucleotide Mix 10 mM (Roche) ; 0,5 µl de Transcriptor Reverse Transcriptase (20 U/µl) (Roche)] ont été ajoutés. Les échantillons ont ensuite été incubés 30 minutes à 55°C, puis 5 minutes à 85°C et finalement 5 minutes sur glace avant d'être placés à -20°C.

13. RT-PCR en temps réel

Les séquences des amorces ont été conçues à l'aide de la version 1.5 du programme Primer Express (Applied Biosystems, Foster City) (Tableau 5).

Tableau 5. *Les sequences des amorces pour la RT-PCR en temps réel.*

Gènes	Amorces	Séquences
TICAM1	Forward	5'-TGCACAGGCCCATCACTTC-3'
	Reverse	5'-AGTTTGTGCTTCAGATACAAGAGCTT-3'
TRAF3	Forward	5'-TCGAAATAATGAATCCAAAATCCTT-3'
	Reverse	5'-TCTCCTTGTCAAGCTCCTTCAGT-3'
TBK1	Forward	5'-CGCACTTTACAGATGAATGTGTTAAA-3'
	Reverse	5'-GCGATAATAACTGTTTCCTAAGATGAAG-3'
IRF3	Forward	5'-AAGGAAGGAGGCGTGTTTGA-3'
	Reverse	5'-CCTTCCGTGAAGGTAATCAGATCT-3'
VAV2	Forward	5'-ACCACACTCAAGTACCCCTACAAGT-3'
	Reverse	5'-GGACTGAGAAAAGAAAAGTTGTAGGAA-3'
RAC2	Forward	5'-AAGCTGGCTCCCATCACCTA-3'
	Reverse	5'-TGAGAGCTGAGCACTCCAGGTA-3'
PAK1	Forward	5'-GGATGAAGGCCAAATTGCA-3'
	Reverse	5'-TGTGAATGACCTGGTTCGAATG-3'
LIMK1	Forward	5'-ATCATCCACCGAGACCTCAACT-3'
	Reverse	5'-GTCAGCCACCACCACATTCTT-3'
CFL2	Forward	5'-CATACGAAACAAAAGAGTCTAAGAAAGAA-3'
	Reverse	5'-CATCTTTAGAGCTAGCATAAATCATCTTG-3'

5 µl d'ADNc dilué 100X ont été mélangés au SYBR Green Master Mix PCR [5 µl d'eau distillée ; 0,84 µl d'amorce Reverse 9 µM ; 0,84 µl d'amorce Forward 9 µM ; 12,5 µl de SYBR green]. Les PCRs ont été réalisées dans un real-time PCR cycler (ABI PRISM 7700 Sequence Detector, PE Applied Biosystems). Les échantillons ont d'abord été incubés 10 minutes à 95°C, puis ont subi 40 cycles de 30 secondes à 95°C, puis ont été incubés 1 minute à 57°C et finalement 30 secondes à 72°C. Les expressions relatives des ARNm d'intérêt ont été comparées selon la méthode des Ct relatifs. La moyenne de l'amplification des ADNc de

23kDa et de l'α-tubuline a été utilisée comme standard endogène afin de normaliser les quantités d'ADNc entre les échantillons.

14. Transfection de siRNA

500.000 cellules MDA-MB-231 ont été cultivées pendant 24 heures dans des boîtes T25. Du milieu de transfection a été préparé : 10 µl de siRNA (20 µM) (ON-TARGET plus SMART pool, Thermo Scientific) ont été ajoutés à 390 µl de milieu Opti-MEM, 8 µl de DharmaFECT ont été ajoutés à 392 µl de milieu Opti-MEM. Après 5 minutes, ces deux solutions ont été mélangées. Après 20 minutes, 3,2 ml de milieu RPMI 1640 (Gibco, Paisley, UK) contenant 10% sérum de veau fœtal ont été ajoutés. Le milieu des boîtes T25 a été remplacé par le milieu de transfection. Après 24 heures, le milieu de transfection a été remplacé par du milieu RPMI 1640 (Gibco, Paisley, UK) contenant 10% sérum de veau fœtal. Finalement, après 24 heures les cellules ont été utilisées pour les différentes expériences.

15. Test de migration

24 heures après la fin de la transfection des siRNA, de petites rayures (scratches) ont été réalisées dans la couche cellulaire des boîtes T25. Les cellules ont ensuite été rincées deux fois avec du PBS. Le milieu cellulaire a été remplacé par du nouveau milieu RPMI 1640 (Gibco, Paisley, UK) contenant 10% sérum de veau fœtal. Des photos des rayures ont été prises tout de suite après qu'elles aient été réalisées, ainsi qu'après 6 heures et 12 heures. Les photos ont été réalisées à l'aide d'un microscope (Leitz) disposant d'un appareil photo (DC-100, Leica). La quantification de la migration cellulaire a été réalisée en comptant le nombre de cellules présentes dans 3 rayures indépendantes.

16. Extraction protéique

Des cellules MDA-MB-231 ont été cultivées dans des boîtes T25. Après avoir enlevé le milieu cellulaire, 200 µl de tampon de lyse (Tris 80 mM pH 7,5, KCl 300 mM, EDTA 2 mM, triton X100 1%), contenant un mélange d'inhibiteurs de protéases (« Complete » de Roche Molecular Biochemicals, 1 tablette dans 2 ml d'eau à une dilution de 1 pour 25) et d'inhibiteurs de phosphatases ($NaVO_3$ 25 mM, PNPP 250 mM, β-glycerophosphate 250 mM et NaF 125 mM à une dilution de 1 pour 25), ont été ajoutés. Les lysats ont été transférés dans

des microtubes et centrifugés 5 minutes à 13.000 tours par minute à 4°C. Les surnageants ont été récoltés et placés à -70°C.

17. Western blot

Les extraits protéiques totaux ont été séparés sur un gel d'électrophorèse SDS-PAGE (sodium dodecyl sulfate-poly-acrylamide gel electrophoresis) et transférés sur une membrane PVDF (polyvinylidene difluoride) (Amersham Biosciences). Le blocage s'est fait dans une boîte contenant 4 ml de LiCor et 4 ml de PBS pendant une heure sous agitation et à température ambiante. La membrane PVDF a ensuite été trempée dans du PBS contenant 0,1% de Tween. La membrane a ensuite été mise en contact avec des anticorps anti-CFL2 (Santa Cruz ; dilué 1 : 100 ; anticorps secondaire de chèvre dilué 1 : 7.500) ou anti-PAK1 (Cell Signalling ; dilué 1 : 1.000 ; anticorps secondaire de lapin dilué 1 : 7.500). La membrane a ensuite été rincée quatre fois 5 minutes dans du PBS contenant 0,1% de Tween, puis deux fois 5 minutes dans du PBS. La membrane a ensuite été séchée pendant 1 heure à 37°C dans l'obscurité. Enfin, la membrane a été scannée avec un scanner Odyssey Infrared Imaging System (LI-COR Biosciences).

18. Test de viabilité cellulaire

Des cellules MDA-MB-231 ont été cultivées à raison de 10.000 cellules par puits dans des plaques 24 puits dans du milieu RPMI 1640 (Gibco, Paisley, UK) contenant 10% sérum de veau fœtal pendant 24, 48, 72 ou 96 heures (avec remplacement du milieu toutes les 24 heures). Après ces incubations, la viabilité des cellules a été évaluée par le test MTT (Sigma, St. Louis, Missouri). 500 µl de solution MTT ont été ajoutés à chaque puits. Les plaques ont été incubées 2 heures à 37°C dans une atmosphère contenant 5% de CO_2. Le mélange de milieu cellulaire et de solution MTT a été enlevé des puits et 1 ml de tampon de lyse [20% SDS (MP Biomedicals, Eschwege, Allemagne), 33,3% N,N-dimethyl-formamide (Merck, Darmstadt, Allemagne), pH 4,7] a été ajouté à chaque puits. Les plaques ont été incubées 1 heure à 37°C dans l'obscurité sous agitation (70 tours par minute). Enfin, les valeurs de densité optique ont été enregistrées à 570 nm (lecteur de plaques, Ultramark Microplate Imaging System, Bio-Rad, Munchen, Allemagne).

III. RESULTATS

1. Résultats *in silico*

1.1. Le choix des CDFs

Comme expliqué dans l'introduction, lors de l'analyse de puces à ADN, l'une des premières étapes est d'établir la correspondance entre les sondes présentes sur les puces et les gènes qu'elles sont censées cibler. Cette correspondance est rendue possible par des fichiers particuliers appelés CDFs. Pour les puces à ADN de la marque Affymetrix, les CDFs standards ont été conçus et distribués par Affymetrix en même temps que les puces elles-mêmes, c'est-à-dire à la fin des années 1990 et au début des années 2000. Depuis lors, les connaissances quant au génome ont évolué. C'est pourquoi certains auteurs ont décidé de réassigner les sondes aux gènes selon des informations plus récentes que celles utilisées par Affymetrix [169, 170]. En particulier, nous avons utilisé, dans le cadre de ce travail, les CDFs alternatifs de l'outil appelé AffyProbeMiner [171].

Afin de montrer que les CDFs d'AffyProbeMiner apportaient une réelle différence par rapport aux CDFs standards d'Affymetrix, nous avons réalisé deux analyses d'un même jeu de données : E-GEOD-1323. Ce jeu de données compare l'expression génique de cellules de cancer du colon métastatiques à celle de cellules non métastatiques [206]. Le seul élément différenciant les deux analyses est le CDF utilisé. Dans les deux cas, le prétraitement des données a été réalisé à l'aide de RMA [175] et le traitement statistique avec le test t de Student [180], comme l'avaient fait les auteurs du jeu de données.

En comparant les résultats des deux analyses, il apparaît que 867 gènes ont une p value d'expression différentielle inférieure à 0,01 lors de l'utilisation du CDF standard d'Affymetrix (Figure 23) alors que nombreux de ces gènes ont une p value supérieure à 0,01 lors de l'utilisation du CDF d'AffyProbeMiner (Figure 24). Inversement, des gènes qui n'étaient pas sélectionnés, lors de l'utilisation du CDF standard d'Affymetrix, le sont quand on utilise le CDF d'AffyProbeMiner.

Ceci montre que l'utilisation de CDFs alternatifs a un impact direct et important sur les résultats. Ils permettent de mettre en évidence d'autres gènes comme exprimés de manière différentielle. Comme les CDFs alternatifs traduisent une information génomique plus récente que les CDFs standards, on peut supposer que les résultats qui en découlent sont plus cohérents avec la réalité biologique. C'est pourquoi, pour le reste des analyses de puces à

ADN réalisées au cours de ce travail, nous n'avons plus utilisé que les CDFs d'AffyProbeMiner.

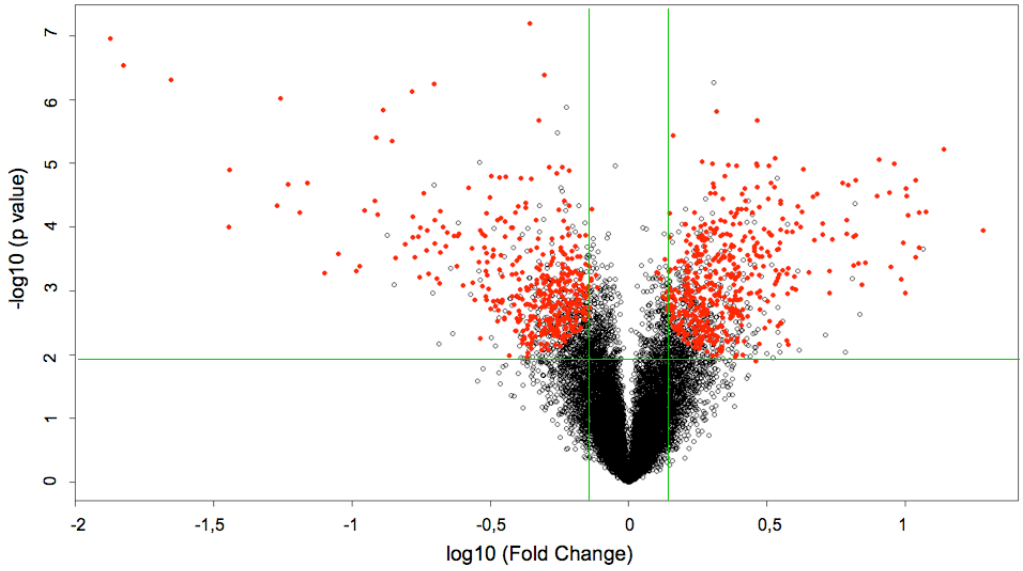


Figure 23. *Volcano plot du jeu de données E-GEOD-1323. En abscisse, se trouve le log10 du fold change de l'expression des gènes. En ordonnée, se trouve le -log10 de la p value de l'expression différentielle des gènes. Un CDF d'Affymetrix a été utilisé, les données ont ensuite été prétraitées avec RMA et traitées avec le test t de Student. Les points rouges représentent les gènes détectés comme exprimés de manière différentielle (p value < 0,01).*

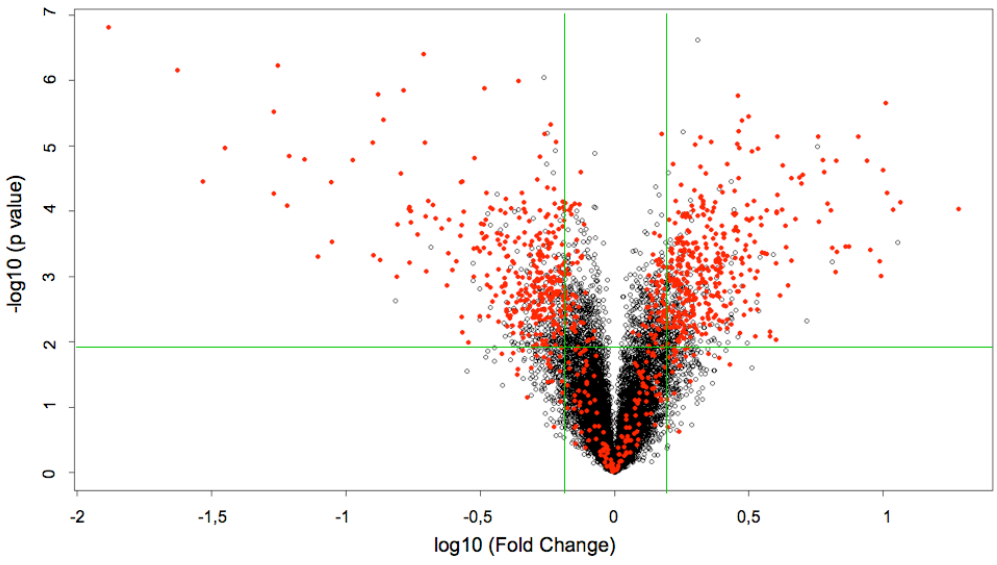


Figure 24. *Volcano plot du jeu de données E-GEOD-1323. En abscisse, se trouve le log10 du fold change de l'expression des gènes. En ordonnée, se trouve le -log10 de la p value de l'expression différentielle des gènes. Un CDF d'AffyProbeMiner a été utilisé, les données ont ensuite été prétraitées avec RMA et traitées avec le test t de Student. Les points rouges représentent les gènes détectés comme exprimés de manière différentielle dans la Figure 23.*

1.2. Méthodologie de sélection de gènes

Dans cette seconde partie des résultats, nous allons présenter, à travers deux articles, la méthodologie de méta-analyse de puces à ADN qui nous a permis d'opérer une sélection des gènes à valider en laboratoire.

Le premier article a été publié le 30 avril 2010 dans la revue BMC Cancer. Son titre est « Meta-analysis of archived DNA microarrays identifies genes regulated by hypoxia and involved in a metastatic phenotype in cancer cells » [207]. Il décrit les 22 jeux de données rapatriés de GEO et ArrayExpress, et comment ils ont été prétraités et traités. Il explique également les trois approches complémentaires qui ont permis de réaliser la sélection des gènes. Il s'agit des intersections, des intersections d'unions et des méta-analyses. Les résultats obtenus sont comparés avec les connaissances trouvées dans la littérature sur les cancers, les métastases et la réponse à l'hypoxie. Une analyse dans DAVID est également menée. Enfin, une série de tests négatifs valident statistiquement les résultats obtenus.

BMC Cancer

RESEARCH ARTICLE Open Access

Meta-analysis of archived DNA microarrays identifies genes regulated by hypoxia and involved in a metastatic phenotype in cancer cells

Michael Pierre[1], Benoît DeHertogh[1], Anthoula Gaigneaux[1], Bertrand DeMeulder[1], Fabrice Berger[1], Eric Bareke[1], Carine Michiels[2] and Eric Depiereux*[1]

Abstract

Background: Metastasis is a major cancer-related cause of death. Recent studies have described metastasis pathways. However, the exact contribution of each pathway remains unclear. Another key feature of a tumor is the presence of hypoxic areas caused by a lack of oxygen at the center of the tumor. Hypoxia leads to the expression of pro-metastatic genes as well as the repression of anti-metastatic genes. As many Affymetrix datasets about metastasis and hypoxia are publicly available and not fully exploited, this study proposes to re-analyze these datasets to extract new information about the metastatic phenotype induced by hypoxia in different cancer cell lines.

Methods: Affymetrix datasets about metastasis and/or hypoxia were downloaded from GEO and ArrayExpress. AffyProbeMiner and GCRMA packages were used for pre-processing and the Window Welch *t* test was used for processing. Three approaches of meta-analysis were eventually used for the selection of genes of interest.

Results: Three complementary approaches were used, that eventually selected 183 genes of interest. Out of these 183 genes, 99, among which the well known *JUNB*, *FOS* and *TP63*, have already been described in the literature to be involved in cancer. Moreover, 39 genes of those, such as *SERPINE1* and *MMP7*, are known to regulate metastasis. Twenty-one genes including *VEGFA* and *ID2* have also been described to be involved in the response to hypoxia. Lastly, DAVID classified those 183 genes in 24 different pathways, among which 8 are directly related to cancer while 5 others are related to proliferation and cell motility. A negative control composed of 183 random genes failed to provide such results. Interestingly, 6 pathways retrieved by DAVID with the 183 genes of interest concern pathogen recognition and phagocytosis.

Conclusion: The proposed methodology was able to find genes actually known to be involved in cancer, metastasis and hypoxia and, thus, we propose that the other genes selected based on the same methodology are of prime interest in the metastatic phenotype induced by hypoxia.

Background

One of the major causes of death by cancer is metastasis. Determining the mechanisms of metastasis initiation and growth should thus improve therapy. Cancer cells have developed many mechanisms to detach from the primary tumor, invade surrounding tissues, migrate and colonize distant organs. These mechanisms include changes in cell-cell and cell-matrix adhesion molecules, extracellular matrix degradation enzymes, cytoskeleton regulation factors and cell-cell communication through cytokines, for example [1]. Recently, high-throughput studies performed in several cancer cell lines identified specific metastasis pathways [2]. However, the exact contribution of these pathways in cell migration and tissue invasion still remains unclear.

Another key feature of a tumor is the presence of hypoxic areas. Hypoxic areas within a tumor are the result of the progressively increasing distance between cells and blood vessels as the tumor is growing, as well as of the abnormal new vasculature. Tumor hypoxia is a

* Correspondence: eric.depiereux@fundp.ac.be
[1] Molecular Biology Research Unit (URBM), University of Namur - FUNDP, Namur, Belgium
Full list of author information is available at the end of the article

marker for poor prognosis. Several hypotheses have been proposed to explain this observation: (i) hypoxia initiates adaptation mediated by the transcription factor HIF-1 which enhances cancer cell survival [3]; (ii) this adaptation also triggers the angiogenesis process; (iii) hypoxia leads to less effective radiotherapy and chemotherapy [4] and (iv) more and more experimental data suggest that hypoxia improves metastasis and/or selects cancer cells with high metastatic potential [5]. Changes in gene expression induced by hypoxia and leading to a migratory and invasive phenotype of tumor cells have been identified. The genes involved code for example for cadherins, plasminogen activators and their receptors, and matrix metalloproteinases [6].

DNA microarrays appeared more than a decade ago and are now one of the most common tools in many molecular biology and medical laboratories. The technique allows the assessment of transcript levels for thousands of genes in one single experiment. Nowadays, thanks to progress in sequencing, an entire genome can be represented on one single microarray [7,8]. In short, a microarray is a physical support to which oligonucleotides (the probes) representing the genes are attached. Then, the labelled mRNAs from a biological sample are incubated with the microarray, thus enabling the labelled mRNAs to hybridize to the probes. The signal is then scanned and several experimental conditions are compared to determine which genes are differentially expressed under a particular condition. Microarrays can differ in the way the probes are attached to the surface, in the length of the probes and in the number of probes per gene, but the principle is always the same.

Affymetrix GeneChips are the most popular type of microarrays. Affymetrix GeneChips display probes that are synthetized *in situ* by photolithography [9]. Typically, each gene in a genome is represented by 11 to 20 perfect match probes of 25 nucleotides and by the same number of mismatch probes differing only in the central nucleotide [10].

Since their release, Affymetrix GeneChips and DNA microarrays in general have encountered many difficulties. The first derives from the fact that if there is only one replicate per condition, no statistical test can be performed at the gene level. Results thus only rely on ratio (fold change) between expression values obtained from different conditions [11]. On the other hand, when several replicates are available, the question can be addressed from a statistical point of view [12]. As one statistical test is performed for each gene, thousands of tests are performed on a single array. This leads to a very large number of false positives (genes detected as differentially expressed when they are not) and false negatives (genes undetected when they are actually differentially expressed), compromising interpretation of the results [13]. When applied, a correction for multiple testing decreases the threshold of significance at such a low level that the number of false negatives increases dramatically. Many statistical methods have recently emerged to solve this problem, from the classical Student *t*-test [14] to more sophisticated tests which fine-tune the estimate of gene variance [15-20]. However, the problem still remains fundamentally unsolved and a large number of expensive replicates are needed to gain in positive and negative predictive power.

Another issue tackled over the past few years concerns the Chip Definition File (CDF), which is involved upstream of the statistical analysis. A CDF is a file developed by Affymetrix that links several probes (probe set) to a given gene name. The probes representing the gene reflect the status of genomic databases several years ago. Since then, genomic information and thus the arguments used to assign given probes to a given probe set have evolved and "alternative" CDFs have emerged [21-23].

Other steps before the statistical processing itself concern pre-processing of the data. Pre-processing steps include background correction within each array of a same experiment, normalization (rescaling) of values between replicates [24-26] and summarization of probe values to obtain a unique expression value per probe set [27]. Scores of methods are available for each pre-processing step [28,29] and a combination of these methods can potentially generate thousands of pre-processings.

A technique's lack of biological reproducibility is not the least of the problems encountered in DNA microarray experiments [30]. Few genes are common to the top gene lists from several experiments, even between subsamples of a same large dataset. Due to the small number of replicates generally used in an experiment and the very large number of genes tested, varying variance estimates lead to differences in the p values associated with any statistical test such that a given gene can randomly move from the first to the thousandth place in the top list.

The fact that there are numerous methods which continue to evolve, combined with the observation of unstable results, constitutes an attractive challenge which can now be tackled thanks to the emergence of public databases such as Gene Expression Omnibus (GEO) [31] and ArrayExpress [32] which collect millions of pieces of expression data. Datasets can be reanalysed from scratch with new parameters (including alternative CDFs) and by combining several datasets relative to the same biological question in one same analysis.

In this study we have tried to find genes, and possibly pathways, that were not previously known to be involved in the metastasis induced by hypoxia. The objective of this paper is to increase our understanding of the disease using new methods to exploit the huge amount of DNA microarrays publicly available. Since the mechanisms

underlying the metastatic phenotype are identical in every cancer cell type [33], this work considered all datasets on metastasis, regardless of the cancer cell type. Our methodology combines well-known published steps from classical analysis with new approaches to analyze several datasets at once. A similar study performed at a smaller scale on breast cancer has been published recently [34], thus validating the feasibility of this type of approach.

Methods

Datasets

All the datasets used in this study were downloaded from two databases (ArrayExpress [32] and Gene Expression Omnibus [31]) and are all generated with Affymetrix platforms. Most of the raw data in .CEL files format were publicly available. If not, the authors were contacted directly. Datasets containing more than one GeneChip model and/or containing more than two conditions were split into sub-datasets. Table 1 presents detailed information about these datasets.

Individual analyses

To process the datasets, we used alternative CDFs from AffyProbeMiner [23]. One CDF is needed per GeneChip model and three packages are needed per CDF. We used Version 1.8.0 of the "CDF distribution" packages. We used Version 1.0.0 of the "PROBE distribution" packages. We used Version 1.1.0 of the "Annotation distribution" packages. The CDFs used were "transcript-consistent", so each probe of a probe set maps to the same set of transcripts. We did not choose "gene-consistent" CDFs (probes of a probe set mapping to transcripts of the same set of genes) to avoid inconsistencies as recommended by Liu *et al*. [23]. The CDFs were chosen based on RefSeq and GeneBank information. The minimal size of a probe set was set to five probes as recommended by Liu *et al*. [23]. Pre-processing was performed with GCRMA [29] with the default parameters. Processing was performed with the Window Welch *t* test [35]. Due to a low number of conditions or of replicates to be statistically useful, datasets GSE4843 and GSE6369 could not be analyzed individually.

These individual analyses provided one gene list for each dataset or sub-dataset. For each gene list, we ranked the genes in ascending order of the p values of their differential expression, such that the most significantly over- or under- expressed genes are located at the top of the list.

Intersections

The results from the individual analyses were grouped into 33 groups. For each group, the 50 most significant genes common to all datasets of the group were selected.

Union intersections

The results from the individual analyses for the 17 metastasis datasets were grouped into 30 groups, while the results from the individual analyses for the 3 hypoxia datasets were grouped in one group. Each metastasis group was considered with the hypoxia group. For each couple of groups, the 50 most significant genes common to at least one dataset of the metastasis group and to at least one dataset of the hypoxia group were selected.

Meta-analyses

The 22 datasets were merged into 14 meta-datasets. Alternative CDFs from AffyProbeMiner [23] were used. The meta-datasets were pre-processed with GCRMA [29] and processed with the Window Welch *t* test [35]. For each meta-dataset, the 50 most significant genes were selected.

Visualization

Genes were thus selected by three approaches: intersections, union intersections and meta-analyses. Some were selected by two or three approaches. Those particular genes were submitted to the webtool DAVID (Database for Annotation, Visualization and Integrated Discovery) [36,37], version 6. The parameters of the "Functional Annotation Tool" were set to retrieve pathway maps from KEGG [38] and Biocarta [39]. And the parameters of the "Functional Annotation Clustering" (a part of the "Functional Annotation Tool") were set to the lowest level of stringency in order to obtain the largest number of maps.

Computer and bioinformatic resources

Individual analyses, intersections, union intersections and meta-analyses were all run with the R statistical software [40] versions 2.4.0 and 2.6.0 and packages from Bioconductor [41] on a 64-bit computer with 4 gb of DDR (biprocessor dual-core Xeon 5160 3.0 Ghz, 8 × 500 gb RAID). Detailed scripts for every approach are provided as additional files (additional files 1, 2, 3, 4 and 5). However, brief descriptions for the individual analyses, intersections, union intersections and meta-analyses are provided here.

For each individual analysis, *expression sets* were obtained with the function `justGCRMA` (with default parameters) from the GCRMA [29] package. The *expression sets* were converted into a matrix with the function `exprs`, then split in two: condition A and condition B. P values were calculated with the Window Welch *t* test [35] with the *pegase* function from the Pegase package. Pegase is a package created by our laboratory that is not yet publicly available. Its function is to process microarray data after pre-processing. It requires an *expression set* as input and returns lists of p values for every well-known processing method as output. Here, *pegase* was run with

Table 1: The datasets retrieved from GEO and ArrayExpress

Data set accession numbers	GeneChip models	Databases	Availability	Experimental conditions
E-GEOD-1323	HG-U133A	AE	Available	3 human colorectal cancer derived from a primary tumor VS. 3 corresponding lymph node metastases
E-GEOD-2280	HG-U133A	AE	Available	8 squamous cell carcinoma of the oral cavity VS. 19 corresponding lymph node metastases
E-MEXP-44	HG-U95Av2	AE	Available	15 head and neck squamous cell carcinoma VS. 3 corresponding lymph node metastases
	HG-UgeneFL	AE	Available	12 head and neck squamous cell carcinoma VS. 11 corresponding lymph node metastases
GSE1056	HG-U95Av2	GEO	Not available	2 human hepatocellular carcinoma under hypoxia for 2 hours VS. 2 control human hepatocellular carcinoma
	HG-U95Av2	GEO	Not available	2 human hepatocellular carcinoma under hypoxia for 24 hours VS. 2 control human hepatocellular carcinoma
GSE2280	HG-U133A	GEO	Available	22 squamous cell carcinoma of the oral cavity VS. 5 corresponding lymph node metastases
GSE2603	HG-U133A	GEO	Available	100 primary breast cancer VS. 21 lung metastases
GSE3325	HG-U133Plus2.0	GEO	Available	7 primary prostate cancer VS. 6 metastases
GSE4086	HG-U133Plus2.0	GEO	Available	2 human Burkitt's lymphoma under hypoxia VS. 2 control human Burkitt's lymphoma
GSE468	HC-G110	GEO	Available	13 primary medulloblastomas VS. 10 metastatic medulloblastomas
GSE4840	HG-U133A	GEO	Not available	3 samples from normal melanocyte culture VS. 12 samples from culture of cutaneous metastasis of melanoma
	HG-U133B	GEO	Not available	3 samples from normal melanocyte culture VS. 12 samples from culture of cutaneous metastasis of melanoma

Table 1: The datasets retrieved from GEO and ArrayExpress (Continued)

GSE4843	HG-U133Plus2.0	GEO	Not available	45 samples from culture of cutaneous melanoma metastasis
GSE6369	HG-U133Plus2.0	GEO	Available	1 primary prostate carcinoma VS. 1 metastatic prostate carcinoma
GSE6919	HG-U95Av2	GEO	Available	65 primary prostate tumors VS. 25 metastatic prostate tumors
	HG-U95B	GEO	Available	66 primary prostate tumors VS. 25 metastatic prostate tumors
	HG-U95C	GEO	Available	65 primary prostate tumors VS. 25 metastatic prostate tumors
GSE7929	HG-U133A	GEO	Available	11 poorly metastatic melanoma VS. 21 highly metastatic melanoma
GSE7930	HG-U133A	GEO	Available	3 poorly metastatic prostate tumors VS. 3 highly metastatic prostate tumors
GSE7956	HG-U133A	GEO	Available	10 poorly metastatic melanoma VS. 29 highly metastatic melanoma
GSE8401	HG-U133A	GEO	Available	31 primary melanoma VS. 52 melanoma metastasis

The GEO or ArrayExpress accession numbers with the corresponding GeneChip model and the experimental conditions.

default parameters. Fold changes were also calculated in the individual analyses.

A *data frame* was built for each dataset. The resulting *data frames* contained the AffyProbeMiner's probe set IDs for every probe set of the chip. They also contained the Entrez Gene IDs [42] corresponding to the probe sets as well as the p values and the fold changes. Since several Entrez Gene IDs can sometimes correspond to the same probe set, these particular probe sets as well as the corresponding p values and fold changes were repeated in the *data frames* with a different Entrez Gene ID each time.

For each *data frame*, the probe sets (and therefore the gene IDs and the fold changes) were ranked in ascending order of the p values of their differential expression. Intersections were created with the `intersect` function applied to the top lists of *data frames* previously described. For groups of datasets where only one GeneChip model was used, intersections were created at the probe set level. But for groups of datasets where several GeneChip models were used, intersections were created at the gene ID level.

Since the group of hypoxia datasets was built with two GeneChip models and was involved in every union intersection, union intersections were all created at the gene ID level. For each union intersection, gene IDs of the top lists of metastasis *data frames* were combined into a vector and gene IDs of the top lists of hypoxia *data frames* were combined into another one. Then, the `intersect` function was applied to those two vectors.

For meta-analyses, *expression sets* were obtained with the `justGCRMA` function (with default parameters). The *expression sets* were converted into a matrix with the `exprs` function, then split in two: condition A and condition B. P values were calculated with the Window Welch *t* test [35] (with default parameters) with the *pegase* function. The probe sets were then ranked in ascending order of the p values of their differential expression, and the 50 most significant ones were selected.

Results and discussion

DNA microarrays and particularly Affymetrix GeneChips are widely used to measure the transcriptome of samples. Since the raw data can now be stored in numeric format, public databases have appeared and reanalysis of archived datasets has become common prac-

tice. Moreover, an increasing number of articles describe analyses combining several datasets. Specific methodologies for such meta-analyses are even published regularly [43-45]. While some of them are large-scale meta-analyses, others target more specific issues especially in the field of oncology. Here, we also provide a strategy for the meta-analysis of specific archived datasets, but the originality of this work is that it combines two different, but intimately related, biological processes: metastasis and hypoxia.

After individual analysis of the datasets retrieved from GEO [31] and ArrayExpress [32] (table 1), the results were combined in an approach called intersections. At each intersection, the 50 most significant genes common to several datasets were selected (figure 1). This arbitrary limit actually represents our upper limit for future *in vitro* validations. Like all statistical thresholds, this figure is arbitrary and is an attempted compromise. Indeed, a higher threshold would generate a number of genes that would be more difficult to interpret. Moreover, a higher threshold would lead to the selection of genes that are not statistically significant in the individual analyses. On the other hand, a lower threshold would allow for selection of a number of genes that would be easier to validate further and more statistically significant, but would also lead to a larger number of false negative genes. Thus, the threshold of 50 genes allows biological interpretation without selecting non significant genes. Since 33 different intersections were designed (Additional file 6), 1650 (33 × 50) different genes could potentially be selected. However, only 704 unique occurrences were obtained because some of them appeared in two or more lists. One advantage of this approach is that it uses the results from different GeneChip models. However, only few genes are common to all GeneChip models. For example, in this study, 8 GeneChip models were used but only 29 genes are represented on all 8 GeneChip models. This is due to some GeneChip models like HC-G110, HG-U133B, HG-U95B and HG-U95C in which few and/or poorly characterized genes are represented. So, intersections combining a large number of GeneChip models need to take a large number of genes into account to obtain the 50 most significant genes common to all the datasets considered.

A second approach combining the results from the individual analyses was the union intersections. By this approach, the 50 most significant genes common to at least one metastasis dataset and to at least one hypoxia dataset were selected at each union intersection. Thirty different union intersections were designed (Additional file 7), each combining a group of metastasis datasets and a group of hypoxia datasets. This approach ensures that every combination of results from the individual analyses takes the hypoxia datasets into account. This step was necessary since fewer hypoxia datasets than metastasis datasets were available. Moreover, union intersections do not require that a large number of genes be taken into account to obtain the 50 most significant genes as fewer are required for a gene to be selected. Out of the 1500 (30 × 50) possible genes, 269 unique occurrences were obtained.

The last approach, called meta-analysis, was not based on the results from the individual analyses. Here, several datasets were merged into single datasets to artificially increase the number of replicates and thus increase the statistical power. Fourteen meta-datasets were designed (Additional file 8). A regular analysis was run on each meta-dataset and the 50 most significant genes were selected (figure 2). Since the meta-analyses were run from scratch, only datasets using the same GeneChip model could be combined. Again, a certain number of genes were present in more than one list: hence, meta-analyses provided 406 different genes out of the 700 (14 × 50) possible genes.

The genes selected by these three approaches are represented in a Venn's diagram (figure 3). The 183 genes selected by more than one approach are considered as the genes of interest. They are listed in the Additional file 9. As six datasets or sub-datasets contain data obtained from melanomas and six others are from the prostate, two supplementary Venn's diagrams have been built based only on these datasets or sub-datasets (additional files 10 and 11). They highlight some genes within the 183 genes of interest. It is interesting to note that, even when half of the datasets are taken into account, so few genes are selected. This shows that the methodology is enriched by the number of datasets and thus by the diversity of information. Interestingly, 99 of the 183 genes of interest are described in the literature to be involved in cancer (figure 4, Additional file 9). For example, the methodology was able to find genes such as *JUNB*, *FOS* and *ATF3*, all members of the AP-1 complex. AP-1 is a transcription factor involved in cell proliferation and differentiation. It has often been described as a "double-edged sword" since its effect can be the repression as well as the promotion of tumorigenesis [46]. Indeed, AP-1 transcription factors are dimers composed of the JUN, FOS and ATF protein families. Depending on the exact AP-1 composition, it promotes or represses tumorigenesis. JUNB acts as a repressor of cell proliferation through its repression activity on the cyclin D1, an essential element in the cell cycle [47,48]. On the contrary, FOS and ATF3 induce oncogenic transformation [49]. Another well-described gene in cancer selected by the methodology is *TP63*. TP63 is a transcription factor sharing a large degree of homology with TP53. It is involved in the development of stratified epithelial tissues [50]. *TP63* has two different promoters and is the target of alternative splicing events leading to the existence of several isoforms.

For example, *TAp63* is a tumor suppressor while Δ*Np63* is an oncogene since it antagonizes *TAp63* [51].

Of the 99 genes known to be involved in cancer, 39 have been described to regulate metastasis (figure 4, Additional file 9). For example, the gene *SERPINE1*, coding for plasminogen activator inhibitor 1 (PAI-1), was selected by the methodology. *SERPINE1* plays a central role in several key steps of metastasis. First, it is able to catalyse degradation of the extracellular matrix to allow penetration of metastatic cancer cells into tissues. Second, when PAI-1 is bound to the plasminogen activator, it is able to modulate cell adhesion by decreasing its affinity for vitronectin and increasing its affinity for endocytic receptors, thus enabling cell migration. Moreover, *SERPINE1* enhances cell proliferation [52]. Another example of a gene selected by the methodology and well described as involved in the metastatic process is the gene coding for matrix metalloproteinase 7 (MMP7). Matrix metalloproteinases are enzymes that cleave the extracellular matrix in normal processes such as morphogenesis, angiogenesis and tissue repair. It was also often described in recent years to be involved in cancer processes such as tumorigenesis, invasion and metastasis [53-55].

Lastly, 21 genes of the 183 selected by the methodology are linked to hypoxia (figure 4, additional file 9). *VEGFA* is probably the best example of such a gene selected by the methodology. *VEGFA* has been largely described to act on endothelial cells to promote the development of vasculature in embryos. Moreover, through the transcription factor HIF-1, hypoxia induces the production of VEGFA to stimulate angiogenesis in newly-formed organs. The same mechanisms are triggered during tumor growth. Indeed, when the tumor size increases, it becomes hypoxic, thus leading to the stabilization of HIF-1 that promotes the transcription of *VEGFA*. VEGFA then stimulates angiogenesis in the tumor [56,57]. *ID2* is another example of genes selected by the methodology and known to be responsive to hypoxia. ID2 belongs to the family of ID proteins which are transcriptional regulators that inhibit basic helix-loop-helix transcription factors in processes such as proliferation, differentiation, development and angiogenesis. It is interesting to note that ID2 is able to inhibit VEGFA and thus limit metastasis [58]. Surprisingly, however, *ID2* is a target of HIF-1 since there are two HIF-1 binding sites within *ID2* gene regulatory sequences. Besides, studies have shown that *ID2* expression is induced under hypoxic conditions [59].

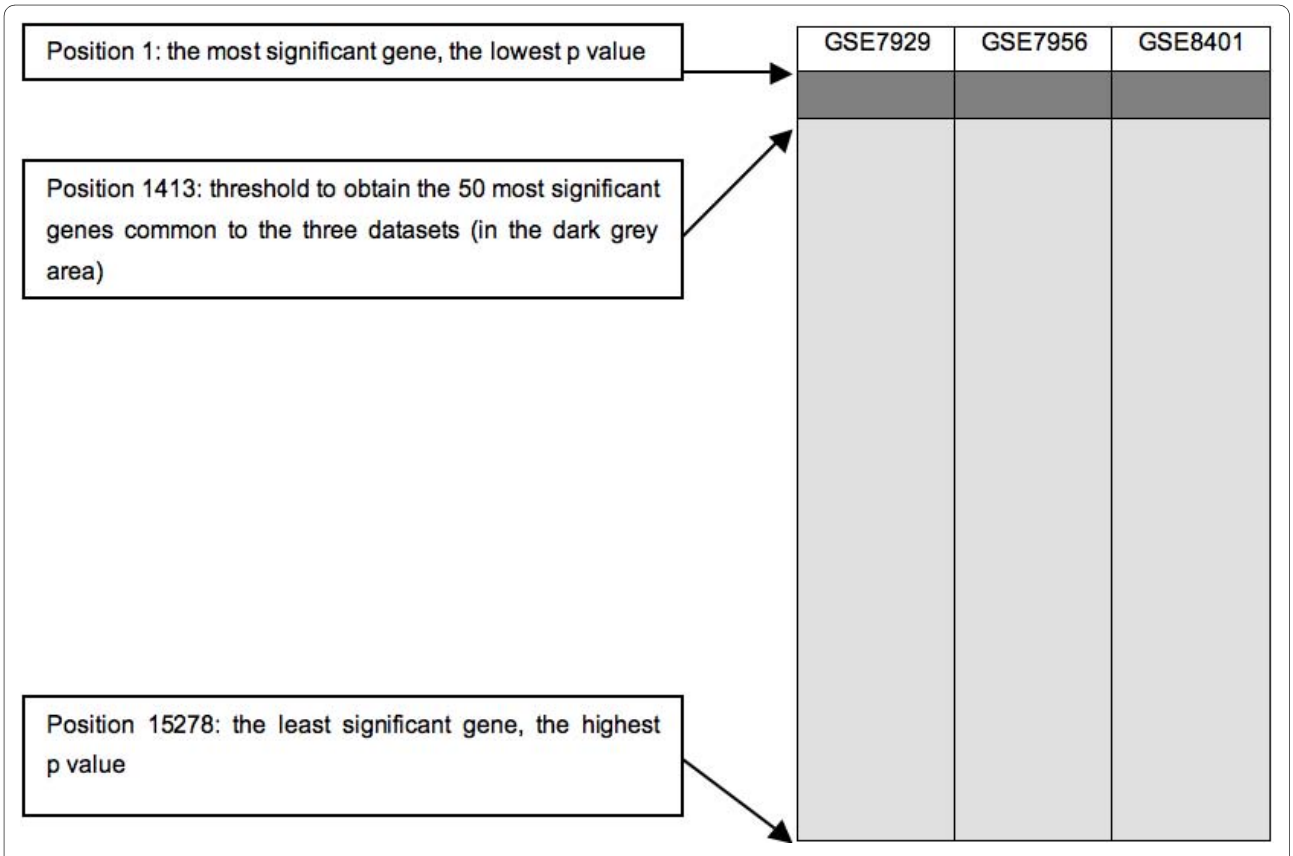


Figure 1 Example of an intersection. In each dataset, the probe sets were ranked in ascending order of the p values. The dark grey area is where the 50 most significant genes common to the three datasets of this particular intersection are found.

Another strong argument in favor of the proposed methodology is the ability of DAVID [36,37] to classify 179 of the 183 genes in 24 different pathways. It is noteworthy that 8 of these pathways are clearly involved in cancer (table 2). For example, "glioma" was one of the pathways retrieved by DAVID [36,37]. Gliomas are cancer which initiate with the oncogenic transformation of a brain or spinal cord cell. There are several types of gliomas which vary in the type of cell transformed. The most common types are those which affect ependymal cells, astrocytes and oligodendrocytes [60]. Another example is "prostate cancer". Prostate cancer is one of the most frequent types of cancer in men. It often develops in patients over the age of 50. This type of cancer is subject to metastasis, particularly in the bones and lymph nodes [61]. "Colorectal cancer" was also retrieved by DAVID [36,37]. This type of cancer is also one of the most common and one of the main cancer-related cause of death. Oncogenic transformation occurs in the adenomatous polyps in the colon and cancer cells can metastasize to the liver, principally [62].

Five other pathways retrieved by DAVID [36,37] are related to proliferation and cell motility (table 2): "focal adhesion", "MAPK signalling pathway", "VEGF signalling pathway", "ErbB signalling pathway" and "regulation of actin cytoskeleton". The focal adhesions are macromolecular structures at the contact points between the cell and the extracellular matrix [63]. They enable tissue remodelling, cell migration and embryogenesis through regulation of the structure of the cytoskeleton, cell adhesion sites and membrane protrusions [64]. The mitogen-activated protein kinase (MAPK) signalling pathway is a cascade involved in the regulation of cellular processes such as cell proliferation, differentiation and stress response [65]. This regulation occurs through the phosphorylation of key proteins in these processes [66]. The VEGF signalling pathway is activated to ensure proliferation and migration of endothelial cells during normal processes such as vasculogenesis as well as pathological processes such as tumor growth [67]. The ErbB signalling pathway is actually composed of several transmembrane receptors able to trigger several signalling pathways when they bind to an extracellular growth factor molecule. These signalling pathways themselves regulate biological processes such as proliferation, differentiation, cell motility and survival [68,69]. The regulation of actin cytoskeleton includes mechanisms which allow for the functions of microfilaments. Microfilaments are responsible for cell shape, intracellular transport and cell motility [70].

As a first negative control, 183 genes were randomly selected. Only 62 of those random genes were found in the literature to be involved in cancer, among which 11 are described to regulate metastasis and 8 are linked to hypoxia. And as a second negative control, 1000 selections of 183 random genes were run. These 1000 lists of random genes were submitted to DAVID to see how

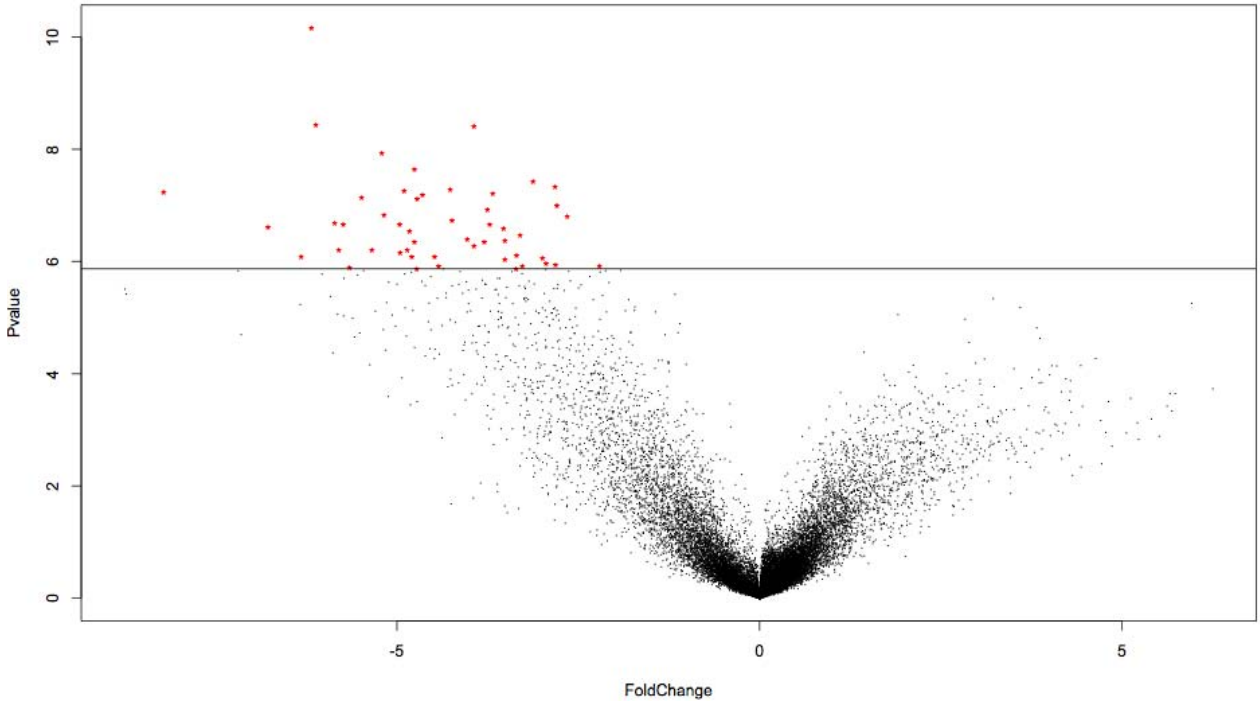


Figure 2 Result of a meta-analysis. The 50 most significant genes were selected in each volcano plot (log2 of the fold changes on the X axis and -log10 of the p values on the Y axis) resulting from the meta-analyses.

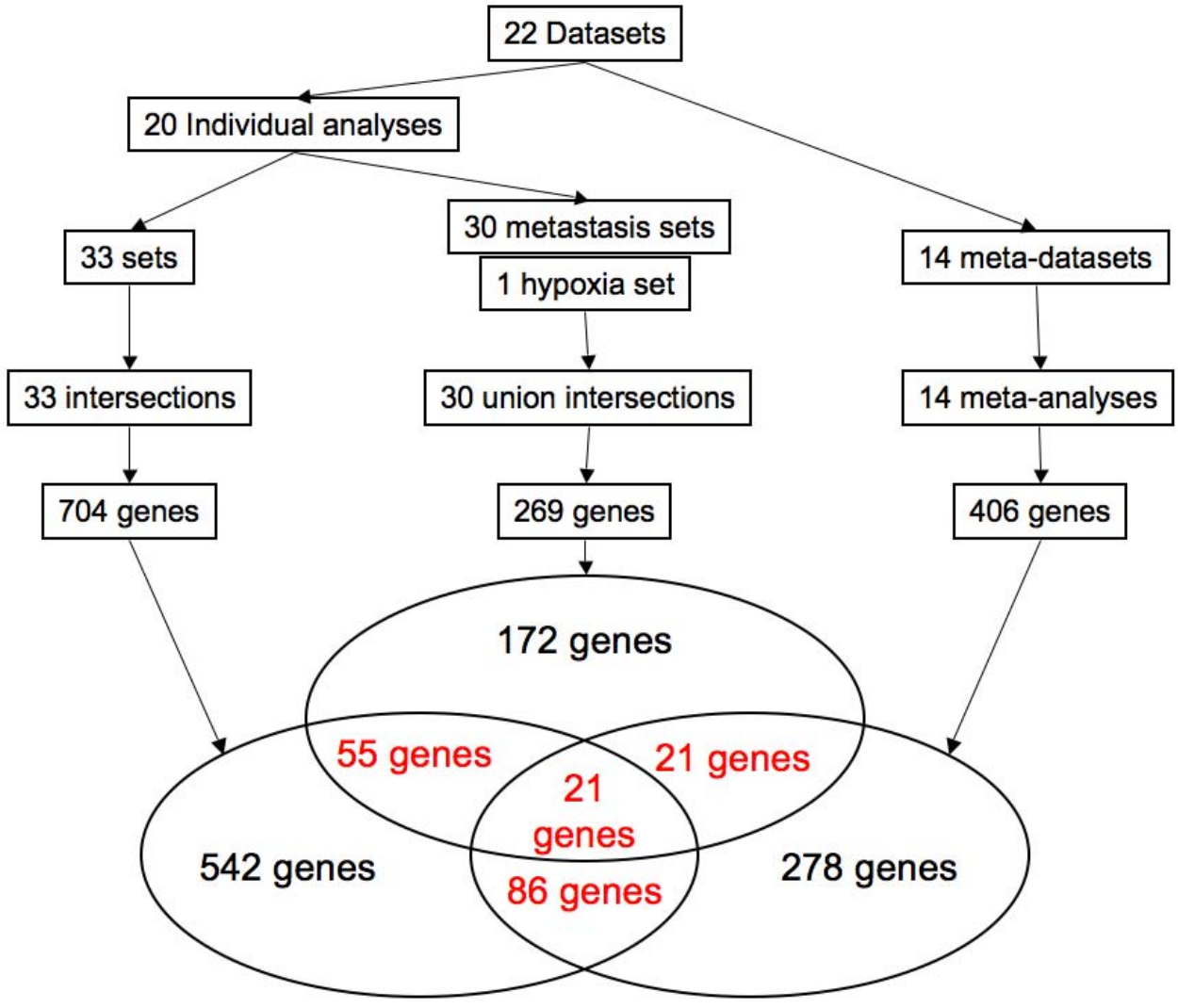


Figure 3 Summary of the methodology. The 22 datasets (or sub-datasets) were used to build several combinations in order to run intersections, union intersections and meta-analyses. These three approaches provided 704, 269 and 406 genes respectively. A Venn's diagram was then generated using this data.

many pathways would be highlighted by chance. This was done for the total number of pathways, for the number of pathways directly involved in cancer and for the number of pathways involved in proliferation and cell motility (figure 5). For the total number of pathways, only two tests gave better results than the 183 genes of interest selected by the methodology. For the number of pathways directly involved in cancer, only nine tests gave equal or better results than the 183 genes of interest selected by the methodology. Lastly, for the number of pathways involved in proliferation and cell motility, only five tests gave equal or better results than the 183 genes of interest selected by the methodology. This indicates that the probability of obtaining the results observed with the 183 genes of interest by chance is between 0,01 and 0,001. Taken together, these results indicate that the methodology is able to find genes actually involved in a particular biological process or even genes involved in a combination of processes (here metastasis and hypoxia).

Since the data on the involvement of the genes of interest in cancer, metastasis and hypoxia support the methodology, we propose that the 84 genes (183 - 99) not known to be involved in cancer to be good candidates for involvement in development of the cancer and in particular in metastasis induced by hypoxia. Obviously, further analyses are required. However, it is already interesting to note that 6 out of the 24 pathways retrieved by DAVID [36,37] concern pathogen recognition and phagocytosis (table 2): "pathogenic *Escherichia coli* infection - EPEC", "pathogenic *Escherichia coli* infection - EHEC", "toll-like receptor signalling pathway", "fMLP induced chemokine gene expression in HMC-1 cells", "Fc epsilon receptor I

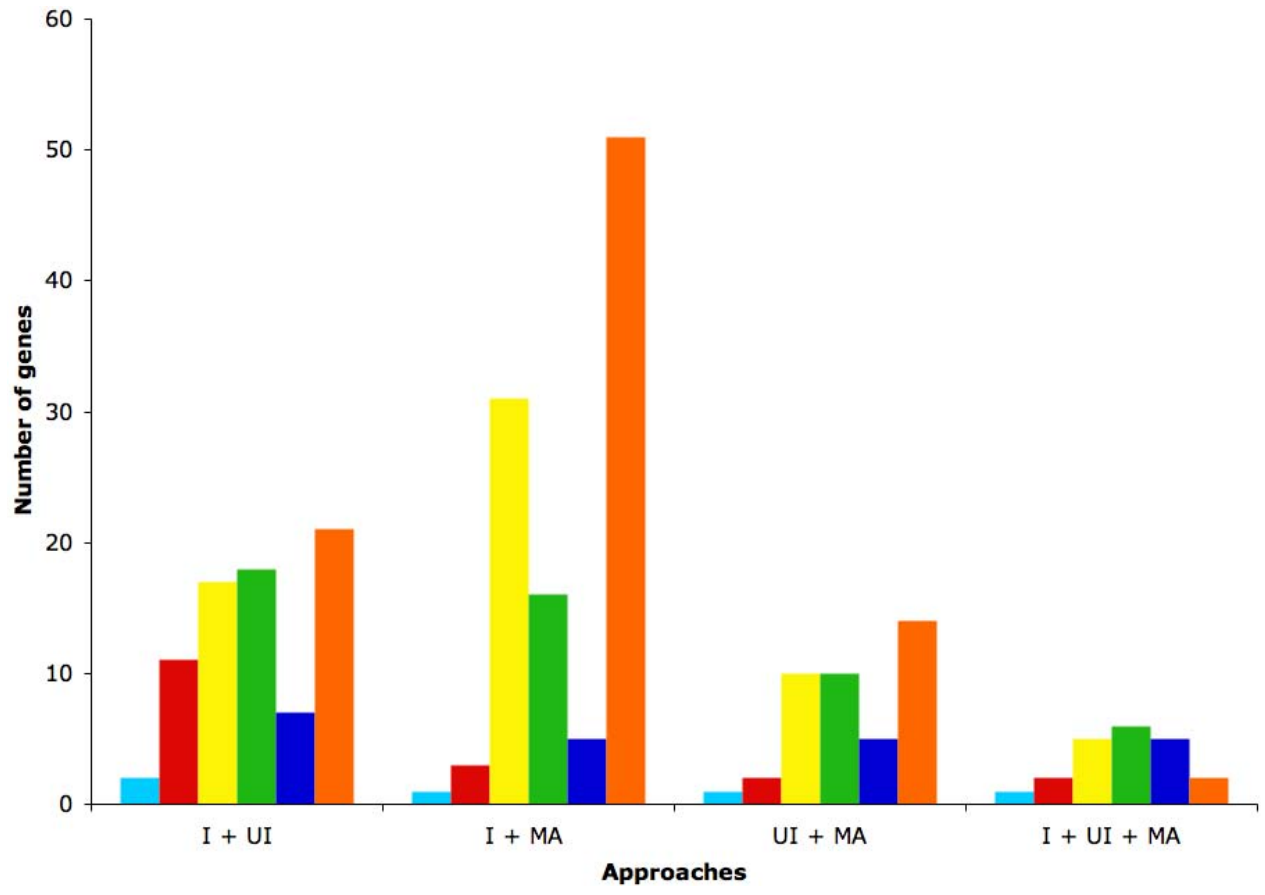


Figure 4 Number of genes involved in processes of interest. After the data mining in the literature, the 183 genes of interest were classified in several categories (light blue: known to be involved in hypoxia, red: known to be involved in cancer and hypoxia, yellow: known to be involved in cancer, green: known to be involved in cancer and metastasis, dark blue: known to be involved in cancer and metastasis and hypoxia, orange: not known to be involved in cancer or metastasis or hypoxia) in function of the combination of approaches (I for intersections, UI for union intersections and MA for meta-analyses).

signalling in mast cells" and "T cell receptor signalling pathway". Thus, we decided to further examine the genes in these pathways and their involvement in cancer, metastasis and hypoxia.

Enteropathogenic *Escherichia coli* (EPEC) and enterohemorrhagic *Escherichia coli* (EHEC) are two pathogens characterized by their ability to cause attaching and effacing lesions. This ability is mainly encoded by the locus of the enterocyte effacement pathogenicity island which includes four genes: *Tir*, *Map*, *EspF* and *EspG*. Interestingly, *Tir* codes for a protein that allows for the accretion of actin [71]. This pathway is thus related to regulation of the actin cytoskeleton pathway identified by the methodology. These three pathways ("pathogenic *Escherichia coli* infection - EPEC", "pathogenic *Escherichia coli* infection - EHEC" and "regulation of actin cytoskeleton") could be activated in metastasis to enable rearrangement of the cytoskeleton and migration of the cell.

The toll-like receptor signalling pathway is composed of a set of receptors able to recognize specific molecules from pathogens. This recognition results in an innate immune response by the activation of inflammatory genes. However, every receptor is specific to a particular signal and triggers a specific cellular response, so the functions of the different toll-like receptors are not redundant [72]. "fMLP-induced chemokine gene expression in HMC-1 cells" is a pathway activated in neutrophils when a bacterial infection occurs. This pathway activates NADPH oxidase that produces reactive oxygen species to kill the bacteria. It also activates genes coding for chemokines to attract other innate immune cells to fight the infection [73]. "Fc Epsilon Receptor I Signalling in Mast Cells" is a defence pathway against some parasites. When activated, mast cells can trigger inflammation [74]. The "T Cell Receptor Signalling Pathway" is a pathway activated when a T Cell Receptor binds to a peptide from a foreign organism. This event activates T cells and immunity [39].

Surprisingly, these pathways have no link with cancer, metastasis or hypoxia, but were identified by our meth-

Table 2: DAVID information

	Pathways	Databases	Genes
Cancer	Prostate cancer	KEGG	MAPK1, IGF1, MAP2K1, CCNE2, NFKBIA
	Chronic myeloid leukemia	KEGG	MAPK1, MAP2K1, NFKBIA
	Colorectal cancer	KEGG	FOS, MAPK1, MAP2K1
	Renal cell carcinoma	KEGG	VEGFA, MAPK1, MAP2K1, PAK6
	Pancreatic cancer	KEGG	STAT1, VEGFA, MAPK1, MAP2K1
	Bladder cancer	KEGG	VEGFA, MAPK1, MAP2K1
	Glioma	KEGG	MAPK1, IGF1, MAP2K1
	Melanoma	KEGG	MAPK1, IGF1, MAP2K1
Proliferation and cell motility	Focal adhesion	KEGG	FLNC, VEGFA, MAPK1, SPP1, IGF1, MAP2K1, PAK6, LAMA3, MYL9
	MAPK signalling pathway	KEGG, BIOCARTA	FLNC, NR4A1, FOS, MAPK1, DUSP1, MAP2K1, DUSP8, NFKBIA
	VEGF signalling pathway	KEGG	VEGFA, HSPB1, MAPK1, MAP2K1
	ErbB signalling pathway	KEGG	MAPK1, MAP2K1, PAK6, ERBB3
	Regulation of actin cytoskeleton	KEGG	ACTG2, MAPK1, MAP2K1, ACTC1, PAK6, MYL9
Pathogen recognition and phagocytosis	Pathogenic Escherichia coli infection - EPEC	KEGG	YWHAZ, TUBB2B, TUBB2A, TUBB2C, TUBB4
	Pathogenic Escherichia coli infection - EHEC	KEGG	YWHAZ, TUBB2B, TUBB2A, TUBB2C, TUBB4
	T Cell Receptor Signalling Pathway	BIOCARTA	FOS, MAP2K1, NFKBIA
	Toll-like receptor signalling pathway	KEGG	STAT1, FOS, MAPK1, SPP1, MAP2K1, NFKBIA
	fMLP induced chemokine gene expression in HMC-1 cells	BIOCARTA	MAPK1, MAP2K1, NFKBIA

Table 2: DAVID information (Continued)

	Fc Epsilon Receptor I Signalling in Mast Cells	BIOCARTA	FOS, MAPK1, MAP2K1
Other	Keratinocyte Differentiation	BIOCARTA	MAPK1, MAP2K1, NFKBIA
	Gap junction	KEGG	TUBB2B, MAPK1, MAP2K1, TUBB2A, TUBB2C, TUBB4
	NFAT and Hypertrophy of the heart	BIOCARTA	MAPK1, IGF1, MAP2K1
	Long-term depression	KEGG	MAPK1, IGF1, MAP2K1
	Cadmium induces DNA synthesis and proliferation in macrophages	BIOCARTA	FOS, MAPK1, MAP2K1, NFKBIA

DAVID classified 179 of the 183 genes of interest into 24 pathways from KEGG or Biocarta. Column 3 presents the genes involved in each specific pathway.

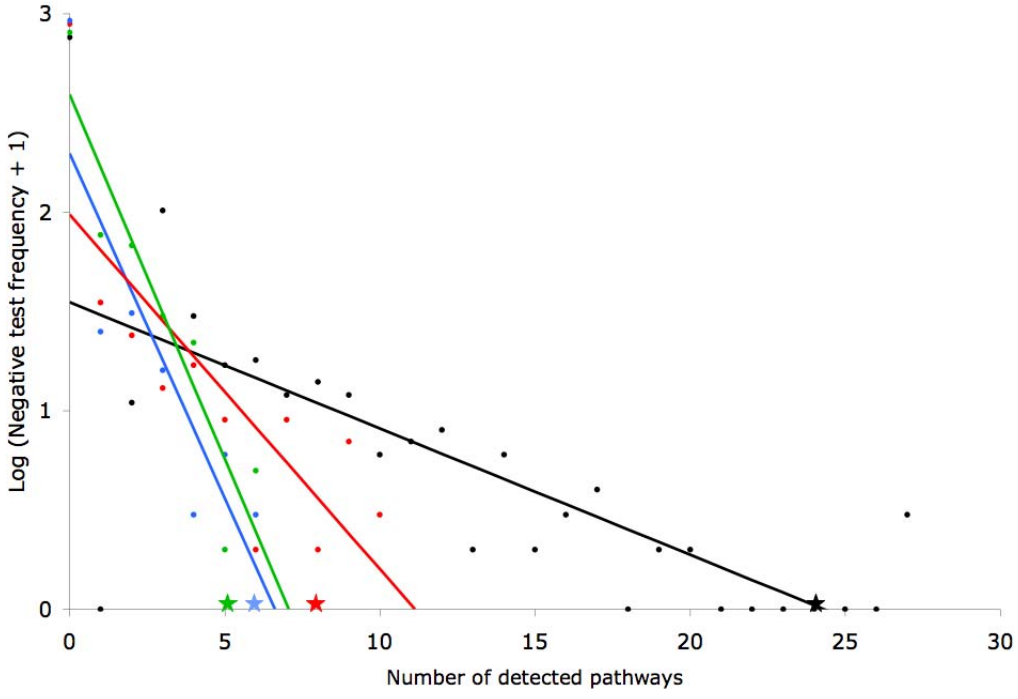


Figure 5 Number of pathways detected by DAVID in negative controls. 1000 lists of 183 random genes were submitted to DAVID. The number of pathways detected per test is presented on the X axis and the logarithm of the frequency of the tests (+ 1 to avoid log (0)) is presented on the Y axis. The black dots show the total number of pathways detected per test and the black star indicates the total number of pathways detected with the 183 genes of interest selected by the methodology. The red dots show the number of pathways directly involved in cancer detected per test and the red star indicates the number of pathways directly involved in cancer detected with the 183 genes of interest selected by the methodology. The green dots show the number of pathways involved in proliferation and cell motility detected per test and the green star indicates the number of pathways involved in proliferation and cell motility detected with the 183 genes of interest selected by the methodology. Lastly, the blue dots show the number of pathways involved in pathogen recognition and phagocytosis detected per test and the blue star indicates the number of pathways involved in pathogen recognition and phagocytosis detected with the 183 genes of interest selected by the methodology.

odology. Moreover, the second negative control assessed the number of pathways involved in pathogen recognition and phagocytosis and only two tests over 1000 trials gave results equal to those with the 183 genes of interest selected by the methodology (figure 5). *In vitro* confirmation of their expression in cancer cell lines with high potential would confirm the relevance of the methodology we propose and the involvement of these genes and pathways in the metastasis of cancer cells. Moreover, functional analysis of the products of these genes should provide new keys to the understanding of the mechanisms involved in the developement of metastases.

Conclusion

We describe a methodology able to identify new genes involved in specific conditions from several microarray datasets. This statement is supported by the fact that this methodology was able to identify genes already known to be involved in the biological processes which we studied. The next step will be *in silico* validation by analysing the expression profile of our genes of interest in publicly available expression profile datasets from different cancer cell lines and *in vitro* validation by qRT-PCR.

The first to be investigated are the genes involved in pathogen recognition and phagocytosis. Indeed, several elements indicate that these pathways may be involved in cancer and particularly in the metastatic process induced by hypoxia. Not only the genes selected by the methodology will be tested. We actually plan to test close neighbouring genes inside the pathways in order to validate large portions of pathways or entire pathways instead of single genes.

This is likely to improve our understanding of the mechanisms underlying this pathology and provide new opportunities to fight it.

Additional material

Additional file 1 R script for individual analyses. The HG-U133A Affymetrix GeneChip was used in this example of script. The script is in R language. Some objects and values, symbolyzed here by X or Y, have to be replace according to the dataset analyzed. CDF packages can vary according to the GeneChip model analyzed.

Additional file 2 R script for data frames. The HG-U133A Affymetrix GeneChip was used in this example of script. The script is in R language. Some objects and values like the length of some vectors have to be replace according to the GeneChip model analyzed. CDF packages can vary according to the GeneChip model analyzed.

Additional file 3 R script for intersections. The script is in R language. Some objects and values, symbolyzed here by X have to be replace according to the datasets involved in the intersection.

Additional file 4 R script for union intersections. The script is in R language. Some objects and values, symbolyzed here by X have to be replace according to the datasets involved in the union intersection.

Additional file 5 R script for meta-analyses. The HG-U133A Affymetrix GeneChip was used in this example of script. The script is in R language. Some objects and values, symbolyzed here by X or Y, have to be replace according to the meta-dataset analyzed. CDF packages can vary according to the GeneChip model analyzed.

Additional file 6 Intersections. 33 groups of datasets were designed based on the experimental conditions and/or the GeneChip model.

Additional file 7 Union intersections. 30 groups of metastasis datasets were designed based on the experimental conditions and/or the GeneChip model. All were compared to the group of hypoxia datasets.

Additional file 8 Meta-datasets. 14 meta-datasets were designed based on the experimental conditions.

Additional file 9 Table of references. This table reports the number of the references in the references section for all 183 genes of interest. These are the publications where those genes were shown to be involved in cancer (column 2), in metastasis (column 3) and/or in hypoxia (column 4).

Additional file 10 Venn's diagram for the prostate datasets. The 6 prostate specific datasets (or sub-datasets) were used to run two intersections, two union intersections and one meta-analysis. These three approaches provided 87, 74 and 48 genes respectively. A Venn's diagram was then generated using these data.

Additional file 11 Venn's diagram for the melanoma datasets. The 6 melanoma specific datasets (or sub-datasets) were used to run three intersections, three union intersections and three meta-analyses. These three approaches provided 144, 97 and 63 genes respectively. A Venn's diagram was then generated using these data.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MP carried out all the experiments, AG and FB participated in the development of the R scripts, BDM participated in the retrieval and the analysis of the informations from DAVID, BDH and EB participated in the selection and the retrieval of the datasets from GEO and ArrayExpress, CM and ED conceived the study, participated in its design and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

M. Pierre is supported by FRIA (Belgium), B. DeMeulder is supported by Televie (Belgium) and C. Michiels is research director of FNRS (Fonds National de la Recherche Scientifique, Belgium).

We thank J.J. LaPres (Biochemistry and Molecular Biology, Michigan State University, East Lansing) for providing the dataset GSE1056 and K.S. Hoek (Department of Dermatology, University Hospital of Zürich, Zürich) for providing the datasets GSE4840 and GSE4843.

Author Details

[1]Molecular Biology Research Unit (URBM), University of Namur - FUNDP, Namur, Belgium and [2]Cell Biology Research Unit (URBC), University of Namur - FUNDP, Namur, Belgium

Received: 30 July 2009 Accepted: 30 April 2010
Published: 30 April 2010

References

1. Friedl P, Wolf K: **Tumour-cell invasion and migration: diversity and escape mechanisms.** *Nat Rev Cancer* 2003, **3(5):**362-374.
2. Pantel K, Brakenhoff RH: **Dissecting the metastatic cascade.** *Nat Rev Cancer* 2004, **4(6):**448-456.
3. Gordan JD, Simon MC: **Hypoxia-inducible factors: central regulators of the tumor phenotype.** *Curr Opin Genet Dev* 2007, **17(1):**71-77.
4. Vaupel P: **The role of hypoxia-induced factors in tumor progression.** *Oncologist* 2004, **9(Suppl 5):**10-17.
5. Sullivan R, Graham CH: **Hypoxia-driven selection of the metastatic phenotype.** *Cancer Metastasis Rev* 2007, **26(2):**319-331.

6. Chan DA, Giaccia AJ: **Hypoxia, gene expression, and metastasis.** *Cancer Metastasis Rev* 2007, **26(2):**333-339.
7. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278(5338):**680-686.
8. Kronick MN: **Creation of the whole human genome microarray.** *Expert Rev Proteomics* 2004, **1(1):**19-28.
9. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D: **Light-directed, spatially addressable parallel chemical synthesis.** *Science* 1991, **251(4995):**767-773.
10. Affymetrix: **Affymetrix Microarray Suite User Guide version 5.0.** Santa Clara: Affymetrix Manual; 2001.
11. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270(5235):**467-470.
12. Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM: **Microarray expression profiling identifies genes with altered expression in HDL-deficient mice.** *Genome Res* 2000, **10(12):**2022-2029.
13. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society* 1995, **57:**289-300.
14. Student: **The Probable Error of a Mean.** *Biometrika* 1908, **6:**1-25.
15. Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes.** *Bioinformatics* 2001, **17(6):**509-519.
16. Cui X, Hwang JT, Qiu J, Blades NJ, Churchill GA: **Improved statistical tests for differential gene expression by shrinking variance components estimates.** *Biostatistics* 2005, **6(1):**59-75.
17. Jain N, Thatte J, Braciale T, Ley K, O'Connell M, Lee JK: **Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays.** *Bioinformatics* 2003, **19(15):**1945-1951.
18. Opgen-Rhein R, Strimmer K: **Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach.** *Stat Appl Genet Mol Biol* 2007, **6:**Article9.
19. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3:**Article3.
20. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98(9):**5116-5121.
21. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, *et al.*: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data.** *Nucleic Acids Res* 2005, **33(20):**e175.
22. Gautier L, Moller M, Friis-Hansen L, Knudsen S: **Alternative mapping of probes to genes for Affymetrix chips.** *BMC Bioinformatics* 2004, **5:**111.
23. Liu H, Zeeberg BR, Qu G, Koru AG, Ferrucci A, Kahn A, Ryan MC, Nuhanovic A, Munson PJ, Reinhold WC, *et al.*: **AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets.** *Bioinformatics* 2007, **23(18):**2385-2390.
24. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19(2):**185-193.
25. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.** *Proc Natl Acad Sci USA* 2001, **98(1):**31-36.
26. Schadt EE, Li C, Ellis B, Wong WH: **Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data.** *J Cell Biochem Suppl* 2001:120-125.
27. Hochreiter S, Clevert DA, Obermayer K: **A new summarization method for Affymetrix probe level data.** *Bioinformatics* 2006, **22(8):**943-949.
28. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4(2):**249-264.
29. Wu Z, Irizarry R, Gentleman R, Murillo F, Spencer F: **A model-based background adjustment for oligonucleotide expression arrays.** *Journal of the American Statistical Association* 2004, **99:**909-917.
30. Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, *et al.*: **Within the fold: assessing differential expression measures and reproducibility in microarray assays.** *Genome Biol* 2002, **3(11):**research0062.
31. **Gene Expression Omnibus** [http://www.ncbi.nlm.nih.gov/geo/]
32. Rocca-Serra P, Brazma A, Parkinson H, Sarkans U, Shojatalab M, Contrino S, Vilo J, Abeygunawardena N, Mukherjee G, Holloway E, *et al.*: **ArrayExpress: a public database of gene expression data at EBI.** *C R Biol* 2003, **326(10-11):**1075-1078.
33. Hunter KW, Crawford NP, Alsarraj J: **Mechanisms of metastasis.** *Breast Cancer Res* 2008, **10(Suppl 1):**S2.
34. Chaudary N, Hill RP: **Hypoxia and metastasis in breast cancer.** *Breast Dis* 2006, **26:**55-64.
35. Berger F, De Hertogh B, Pierre M, Gaigneaux A, Depiereux E: **The "Window t test": a simple and powerful approach to detect differentially expressed genes in microarray datasets.** *Central European Journal of Biology* 2008, **3(3):**327-344.
36. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4(5):**P3.
37. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4(1):**44-57.
38. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 1999, **27(1):**29-34.
39. **Biocarta Pathways** [http://www.biocarta.com/genes/index.asp]
40. Ihaka R, Gentleman R: **R: a language for data analysis and graphics.** *Journal of Computational and Graphical Statistics* 1996, **5:**299-314.
41. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, *et al.*: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5(10):**R80.
42. **Entrez Global Query Cross-Database Search System** [http://www.ncbi.nlm.nih.gov/sites/gquery]
43. Gur-Dedeoglu B, Konu O, Kir S, Ozturk AR, Bozkurt B, Ergul G, Yulug IG: **A resampling-based meta-analysis for detection of differential gene expression in breast cancer.** *BMC Cancer* 2008, **8:**396.
44. Ma S, Huang J: **Regularized gene selection in cancer microarray meta-analysis.** *BMC Bioinformatics* 2009, **10:**1.
45. Ochsner SA, Steffen DL, Hilsenbeck SG, Chen ES, Watkins C, McKenna NJ: **GEMS (Gene Expression MetaSignatures), a Web resource for querying meta-analysis of expression microarray datasets: 17beta-estradiol in MCF-7 cells.** *Cancer Res* 2009, **69(1):**23-26.
46. Eferl R, Wagner EF: **AP-1: a double-edged sword in tumorigenesis.** *Nat Rev Cancer* 2003, **3(11):**859-868.
47. Jochum W, Passegue E, Wagner EF: **AP-1 in mouse development and tumorigenesis.** *Oncogene* 2001, **20(19):**2401-2412.
48. Shaulian E, Karin M: **AP-1 in cell proliferation and survival.** *Oncogene* 2001, **20(19):**2390-2400.
49. van Dam H, Castellazzi M: **Distinct roles of Jun: Fos and Jun: ATF dimers in oncogenesis.** *Oncogene* 2001, **20(19):**2453-2464.
50. Tomkova K, Tomka M, Zajac V: **Contribution of p53, p63, and p73 to the developmental diseases and cancer.** *Neoplasma* 2008, **55(3):**177-181.
51. Malaguarnera R, Vella V, Vigneri R, Frasca F: **p53 family proteins in thyroid cancer.** *Endocr Relat Cancer* 2007, **14(1):**43-60.
52. Fabre-Guillevin E, Malo M, Cartier-Michaud A, Peinado H, Moreno-Bueno G, Vallee B, Lawrence DA, Palacios J, Cano A, Barlovatz-Meimon G, *et al.*: **PAI-1 and functional blockade of SNAI1 in breast cancer cell migration.** *Breast Cancer Res* 2008, **10(6):**R100.
53. Beeghly-Fadiel A, Shu XO, Long J, Li C, Cai Q, Cai H, Gao YT, Zheng W: **Genetic polymorphisms in the MMP-7 gene and breast cancer survival.** *Int J Cancer* 2009, **124(1):**208-214.
54. Fang YJ, Lu ZH, Wang GQ, Pan ZZ, Zhou ZW, Yun JP, Zhang MF, Wan DS: **Elevated expressions of MMP7, TROP2, and survivin are associated with survival, disease recurrence, and liver metastasis of colon cancer.** *Int J Colorectal Dis* 2009, **24(8):**875-884.
55. Liu D, Nakano J, Ishikawa S, Yokomise H, Ueno M, Kadota K, Urushihara M, Huang CL: **Overexpression of matrix metalloproteinase-7 (MMP-7) correlates with tumor proliferation, and a poor prognosis in non-small cell lung cancer.** *Lung Cancer* 2007, **58(3):**384-391.
56. Neufeld G, Cohen T, Gengrinovitch S, Poltorak Z: **Vascular endothelial growth factor (VEGF) and its receptors.** *Faseb J* 1999, **13(1):**9-22.
57. Roskoski R Jr: **Vascular endothelial growth factor (VEGF) signaling in tumor progression.** *Crit Rev Oncol Hematol* 2007, **62(3):**179-213.
58. Tsunedomi R, Iizuka N, Tamesa T, Sakamoto K, Hamaguchi T, Somura H, Yamada M, Oka M: **Decreased ID2 promotes metastatic potentials of**

hepatocellular carcinoma by altering secretion of vascular endothelial growth factor. *Clin Cancer Res* 2008, **14(4):**1025-1031.
59. Lofstedt T, Jogi A, Sigvardsson M, Gradin K, Poellinger L, Pahlman S, Axelson H: **Induction of ID2 expression by hypoxia-inducible factor-1: a role in dedifferentiation of hypoxic neuroblastoma cells.** *J Biol Chem* 2004, **279(38):**39223-39231.
60. Chandana SR, Movva S, Arora M, Singh T: **Primary brain tumors in adults.** *Am Fam Physician* 2008, **77(10):**1423-1430.
61. Kaliks RA, Del Giglio A: **Management of advanced prostate cancer.** *Rev Assoc Med Bras* 2008, **54(2):**178-182.
62. Alberts SR: **Updated options for liver-limited metastatic colorectal cancer.** *Clin Colorectal Cancer* 2008, **7(Suppl 2):**S58-62.
63. Petit V, Thiery JP: **Focal adhesions: structure and dynamics.** *Biol Cell* 2000, **92(7):**477-494.
64. Mitra SK, Hanson DA, Schlaepfer DD: **Focal adhesion kinase: in command and control of cell motility.** *Nat Rev Mol Cell Biol* 2005, **6(1):**56-68.
65. Tanoue T, Nishida E: **Docking interactions in the mitogen-activated protein kinase cascades.** *Pharmacol Ther* 2002, **93(2-3):**193-202.
66. Biondi RM, Nebreda AR: **Signalling specificity of Ser/Thr protein kinases through docking-site-mediated interactions.** *Biochem J* 2003, **372(Pt 1):**1-13.
67. Hoeben A, Landuyt B, Highley MS, Wildiers H, Van Oosterom AT, De Bruijn EA: **Vascular endothelial growth factor and angiogenesis.** *Pharmacol Rev* 2004, **56(4):**549-580.
68. Holbro T, Hynes NE: **ErbB receptors: directing key signaling networks throughout life.** *Annu Rev Pharmacol Toxicol* 2004, **44:**195-217.
69. Yarden Y, Sliwkowski MX: **Untangling the ErbB signalling network.** *Nat Rev Mol Cell Biol* 2001, **2(2):**127-137.
70. Pollard TD: **The cytoskeleton, cellular motility and the reductionist agenda.** *Nature* 2003, **422(6933):**741-745.
71. Kaper JB, Nataro JP, Mobley HL: **Pathogenic Escherichia coli.** *Nat Rev Microbiol* 2004, **2(2):**123-140.
72. Kawai T, Akira S: **Antiviral signaling through pattern recognition receptors.** *J Biochem* 2007, **141(2):**137-145.
73. Dewas C, Fay M, Gougerot-Pocidalo MA, El-Benna J: **The mitogen-activated protein kinase extracellular signal-regulated kinase 1/2 pathway is involved in formyl-methionyl-leucyl-phenylalanine-induced p47phox phosphorylation in human neutrophils.** *J Immunol* 2000, **165(9):**5238-5244.
74. Kitaura J, Xiao W, Maeda-Yamamoto M, Kawakami Y, Lowell CA, Kawakami T: **Early divergence of Fc epsilon receptor I signals for receptor up-regulation and internalization from degranulation, cytokine production, and survival.** *J Immunol* 2004, **173(7):**4317-4323.

Pre-publication history

The pre-publication history for this paper can be accessed here:
http://www.biomedcentral.com/1471-2407/10/176/prepub

doi: 10.1186/1471-2407-10-176
Cite this article as: Pierre *et al.*, Meta-analysis of archived DNA microarrays identifies genes regulated by hypoxia and involved in a metastatic phenotype in cancer cells *BMC Cancer* 2010, **10**:176

Le second article a été publié le 17 février 2011 dans la revue Journal Of Proteomics And Bioinformatics. Son titre est « Enhanced Meta-analysis Highlights Genes Involved in Metastasis from Several Microarray Datasets » [208]. Dans cet article, nous avons apporté une modification dans l'une des approches permettant la sélection des gènes. En effet, précédemment, les intersections demandaient de fixer un seuil arbitraire pour sélectionner les gènes des différents jeux de données. Ici, après avoir consulté plusieurs biostatisticiens, nous avons développé une méthode statistique permettant de calculer avec précision le nombre de gènes à sélectionner pour que le résultat soit statistiquement valide. Il en est résulté une nouvelle série de gènes candidats pour la validation en laboratoire. En plus d'être comparés à la littérature, analysés dans DAVID et validés par des tests négatifs, ces gènes ont été validés par des profils d'expression réalisés à partir de jeux de données indépendants.

De cette méta-analyse de puces à ADN, 165 gènes ont été sélectionnés. Parmi ceux-ci, 91 sont déjà connus pour être impliqués dans le cancer, dont 41 dans le phénotype métastatique, et 20 dans la réponse à l'hypoxie. De plus, ces 165 gènes ont été replacés dans 42 voies de signalisation par DAVID. Parmi celles-ci, 12 sont directement liées au cancer et 5 à la prolifération et à la mobilité cellulaire. De plus, les tests négatifs réalisés sur des gènes sélectionnés aléatoirement ont montré que la probabilité d'obtenir de tels résultats était inférieure à 1%.

Nous en concluons donc que cette méthodologie permet une méta-analyse originale de puces à ADN. En effet, elle est capable de mettre en évidence des gènes impliqués dans un ou plusieurs processus biologiques d'intérêt. Parmi ceux-ci, certains sont déjà connus dans la littérature pour être impliqués dans ces processus, validant ainsi la méthodologie. D'autres, par contre, n'ont pas encore été répertoriés comme participant à ces phénomènes. Ce sont ces gènes qui présentent les plus belles perspectives de recherches futures.

Dans le cas présent, ce sont les gènes représentant cinq voies de signalisation impliquées dans la reconnaissance de pathogènes et la phagocytose qui ont particulièrement attiré notre attention et auxquels nous avons consacré les validations *in vitro* pour mettre en évidence leur éventuel rôle dans la migration des cellules cancéreuses.

Proteomics & Bioinformatics

Pierre et al. J Proteomics Bioinform 2011, 4:1
http://dx.doi.org/10.4172/jpb.1000164

Research Article Open Access

Enhanced Meta-analysis Highlights Genes Involved in Metastasis from Several Microarray Datasets

Michael Pierre[1], Benoît DeHertogh[1], Bertrand DeMeulder[1], Eric Bareke[1], Sophie Depiereux[1], Carine Michiels[2] and Eric Depiereux[1]*

[1]Molecular Biology Research Unit (URBM), University of Namur - FUNDP, Namur, Belgium
[2]Cell Biology Research Unit (URBC), NARILIS, University of Namur - FUNDP, Namur, Belgium

Abstract

Metastasis is the final stage of cancer and is still associated with high mortality despite breakthroughs in recent years. Hypoxia at the center of the primary tumor is a major cause of metastasis. Here, we present a new meta-analysis-based methodology to pick out genes involved in one or two biological processes from several microarray datasets using a statistic that avoids the definition of an arbitrary threshold, providing statistically-significant results. Applied to metastasis and hypoxia datasets, this methodology was able to select genes already known to be involved in these phenomena as well as new candidates for further analyses.

165 genes of interest were selected, many of which were already known to be involved in cancer, metastasis and/or hypoxia. Moreover, some could be classified into 42 pathways, including 12 cancer pathways and 5 proliferation and cell motility pathways. Negative tests performed with random genes failed to provide such results. In additional independent validations, expression profiles were generated for the 165 genes of interest from two other datasets with MDA-MB-231, MCF-7 and L3.6pl cells and the previous results were confirmed in most cases.

Keywords: Metastasis; Hypoxia; Microarray; Meta-analysis; Statistical threshold

Abbreviations: ADM : Adrenomedullin; AFP: Alpha-Fetoprotein; ASPM: asp (abnormal spindle) homolog, microcephaly associated (Drosophila); BCAT1: Branched Chain Amino-acid Transaminase 1, cytosolic; BCL-2: B-cell CLL/lymphoma 2; BHLHE40: Basic Helix-Loop-Helix Family, Member e40; CAV1: Caveolin 1, caveolae protein, 22kDa; CDF: Chip Definition File; DAVID: Database for Annotation, Visualization and Integrated Discovery; GCRMA: GeneChip Robust Multi-array Analysis; GEO: Gene Expression Omnibus; GLUT3, glucose transporter type 3; GPC1: glypican 1; GSE: GEO series; HBGF: Heparin-Binding Growth Factor; HMC-1: Human Mast Cell-1; KEGG: Kyoto Encyclopedia of Genes and Genomes; MAP2K1: Mitogen-Activated Protein Kinase Kinase 1; MMP: Matrix Metalloproteinase; NR4A1: Nuclear Receptor subfamily 4, group A, member 1; RGS20: Regulator of G-protein Signaling 20; SLC2A3: Solute carrier family 2 (facilitated glucose transporter), member 3; SMURF2 : SMAD specific E3 Ubiquitin protein ligase 2; TNF-β: Tumor Necrosis Factor; ZFHX3: zinc finger homeobox 3

Introduction

Metastasis is the final stage of cancer and is characterized by the migration of primary tumor cells to distant organs [1]. These cells set up various mechanisms in a sequential fashion [2]. First, they lose their adherence to the other tumor cells and gain adherence to the extracellular matrix. Then, they degrade the extracellular matrix to invade the tissue. Next, they enter blood or lymph vessels and circulate in the body until they leave the bloodstream or lymphatic circulation to divide in the organ where they stop. These mechanisms involve changes in expression profiles of genes such as integrins, matrix metalloproteinases and growth factors.

Hypoxia within the primary tumor further enhances the metastatic phenotype. Hypoxia occurs at the center of the tumor because the distance between the cells and blood vessels increases as a result of tumor growth and because the new vasculature is abnormal [3]. Hypoxia selects cancer cells with a high metastatic potential [4] and triggers survival mechanisms, leading to increased radiotherapy and chemotherapy resistance [5].

Developed during the 1990s, DNA microarrays are used in an increasing number of applications in molecular biology research. Despite the technique's ability to assess the entire transcriptome of an organism at once [6], it is associated with many difficulties in the analysis of results. Several issues can be pointed out. First, to produce statistical results, several replicates are needed. However, since several thousands of tests are performed at once, the number of false positives and false negatives rapidly becomes unmanageable [7,8]. Therefore, the only solution is to increase the number of replicates, but the cost of the analyses prevents this. Second, the chip probes do not always correspond to the genes they are expected to find, which requires regular updating of the files linking probes and genes [9,10]. Finally, the number of possible combinations of analysis methods frequently leads to inappropriate choices.

These problems and the growing number of publicly-available datasets have led the research community to try new ways to analyze DNA microarrays. Meta-analysis is one of these solutions. It consists of analyzing several related datasets at once [11-13]. This work proposes a new approach to set a statistically-significant threshold to achieve a more relevant meta-analysis. This new methodology was applied to metastasis and hypoxia datasets and the results were validated in an independent experiment in which another dataset assessing MDA-MB-231 and MCF-7 cells was used to generate expression profiles for each gene selected by the methodology. Since MDA-MB-231 cells are breast cancer cells with high metastatic potential and MCF-7 cells are

***Corresponding author:** Eric Depiereux, Molecular Biology Research Unit (URBM), University of Namur, 61 rue de Bruxelles, 5000 Namur, Belgium, Tel: +32 81 72 44 15; Fax: +32 81 72 44 20; E-mail: eric.depiereux@fundp.ac.be

Received November 28, 2010; **Accepted** February 14, 2011; **Published** February 17, 2011

Citation: Pierre M, DeHertogh B, DeMeulder B, Bareke E, Depiereux S, et al. (2011) Enhanced Meta-analysis Highlights Genes Involved in Metastasis from Several Microarray Datasets. J Proteomics Bioinform 4: 036-043. doi:10.4172/jpb.1000164

breast cancer cells with poor metastatic potential, these expression profiles validate the involvement of the genes of interest in the cell motility process. As metastasis mechanisms are the same in every type of cancer [14], this meta-analysis was run independently of the type of cancer. Certain genes involved in hypoxia were also validated by expression profiles in another independent experiment in which highly metastatic pancreatic cancer cells (L3.6pl) were exposed to normoxia or hypoxia. These results may help us to discover new targets to fight metastasis, and particularly in its upregulation by hypoxia.

Materials and Methods

Datasets, individual analyses, union intersections and meta-analyses

All the datasets and procedures used to run the individual analyses, union intersections and meta-analyses were described in Pierre et al. [15]. The data was pre-processed with GCRMA (GeneChip Robust Multi-array Analysis) [16] and the Window Welch *t* test [17] was used for the processing, according to the benchmark performed by De Hertogh et al. [18]. The additional datasets, GSE (GEO series) 5823 and GSE9350, used for determining expression profiles were downloaded from GEO (Gene Expression Omnibus) (NCBI, 2000).

Intersections

33 groups of individual analyses were designed as described in Pierre et al. [15]. A threshold rank was calculated in each group with the equation (1):

$$r = [1-(1-P)^{1/n}]^{1/k} \times N \quad (1)$$

where r = the threshold rank, P = the fixed probability, n = the number of genes suspected to be involved in metastasis and/or in the response to hypoxia, k = the number of datasets in the group and N = the number of probe sets on the GeneChip (the largest when several GeneChip models are involved in the group). This equation is explained in the discussion.

The genes common to all datasets of the group and above the threshold were selected in each group.

Visualization

The webtool DAVID (Database for Annotation, Visualization and Integrated Discovery) [19,20], version 6.7, was used to visualize the genes of interest on KEGG (Kyoto Encyclopedia of Genes and Genomes) [21] and Biocarta (Biocarta) pathway maps. The largest number of maps was obtained by setting the stringency of the "Functional Annotation Clustering" to the lowest level.

Expression profiles

The datasets GSE5823 and GSE9350 were analyzed separately with an AffyProbeMiner's CDF (chip definition file) [10] and pre-processed with GCRMA [16] with default parameters. The expression values of each probe set for each gene of interest in the dataset GSE5823 were then plotted for GeneChips where control MCF-7 cells or control MDA-MB-231 cells were analyzed. The expression values of genes known to be involved in hypoxia in the dataset GSE9350 were also plotted for GeneChips where L3.6pl cells were exposed to hypoxia or normoxia.

Computer and bioinformatics resources

Versions 2.4.0, 2.6.0 and 2.10.1 of the R statistical software [22] and the Bioconductor [23] and AffyProbeMiner [10] packages were used on a 64-bit computer with 4gb of DDR (biprocessor dual-core Xeon 5160 3.0Ghz, 8 x 500gb RAID).

Results

Intersections

An intersection is composed of a group of datasets. For each of these datasets, the genes are ranked in ascending order of the p values of their differential expression. The intersection approach involves selecting the genes that are common to all the top lists of these datasets. To do this, the top lists must be defined and a maximal rank must be considered. A statistic was developed in order to calculate this rank to ensure that all selected genes of the intersection are selected with statistical significance [24]. This statistic takes into account the number of probe sets of the GeneChip, the number of genes potentially involved in metastasis and/or the response to hypoxia and the number of datasets involved in the intersection. Figure 1 presents the logarithm of the threshold ranks + 1 (to avoid log (0)) as well as the logarithm of the number of selected genes + 1 (to avoid log (0)) for the 33 intersections. The 33 intersections selected 2,656 genes, among which 846 were non redundant. The number of genes selected by intersection varied between 0 and 513.

Union intersections

Since the 846 genes selected by the intersection approach are too many genes to further process, two other approaches were added. The first is the union intersection approach. Each union intersection takes the hypoxia datasets into account, comparing the group of hypoxia datasets to a group of metastasis datasets and selecting the 50 most significant genes common to at least one hypoxia dataset and to at least one metastasis dataset. Here, no statistic exists to set a threshold as in the intersection approach. Hence, an arbitrary threshold of 50 genes was set. However, unlike intersections, union intersections do not require a large maximal rank to select 50 genes as less is required for a gene to be selected. Thirty union intersections were designed and 1,500 genes (30 x 50) were selected by this approach, among which 269 are unique occurrences.

Meta-analyses

The last approach used to reduce the number of genes to be considered is the meta-analysis approach. This approach is not based on the results of individual analyses. Here, several datasets are merged

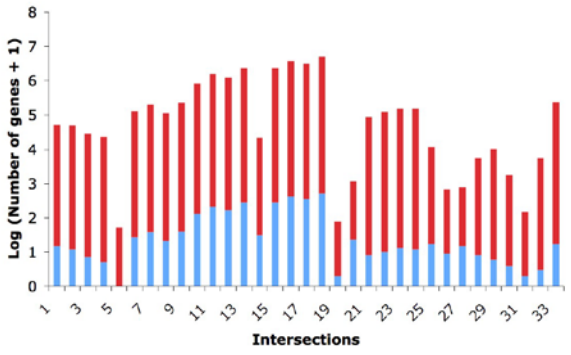


Figure 1: Threshold ranks and number of genes selected by intersections. A threshold rank was calculated for each intersection with the equation $r = [1-(1-p)^{1/n}]^{1/k} \times N$ where p represents the probability a gene has to be selected, r the maximum rank of a gene to be selected with the probability p, N the number of genes represented on the microarray, k the number of datasets taken into account in the intersection and n the number of genes likely to be involved in the phenomenon studied. The red bars show the logarithm of the threshold ranks + 1 (to avoid log (0)). The threshold ranks selected a defined number of genes per intersection. The blue bars show the logarithm of the number of selected genes + 1 (to avoid log (0)).

into meta-datasets. Then, classical analyses are run on these meta-datasets and the 50 most significant genes are selected. Again, no statistical threshold was set, first of all because the 50 selected genes are significant anyway and secondly because setting a threshold such as 0.05 would produce too many selected genes for the meta-analyses. Fourteen meta-datasets were designed and 700 (14 x 50) genes were selected by the meta-analysis approach, of which 406 were unique occurrences.

Combination of approaches

To select an appropriate number of genes to further validate, the three approaches (intersection, union intersection and meta-analysis) were combined. Fifteen genes were found to be common to the three approaches, 52 genes were only common to the intersections and union intersections, 71 genes were only common to the intersections and meta-analyses and 27 genes were only common to the union intersections and meta-analyses. These 165 genes were considered as genes of interest and are highlighted in a Venn's diagram (Figure 2). Among these 165 genes of interest, 91 are already known in the literature to be involved in cancer, 41 in metastasis and 20 in the response to hypoxia (additional file 1).

Visualization

The 165 genes of interest were submitted to DAVID [19,20] to retrieve pathway maps from KEGG [21] and Biocarta (Biocarta). Forty-two different pathways were retrieved, among which 12 are directly involved in cancer and 5 are known to be involved in cell proliferation and cell motility (Table 1). To ensure that this result was not simply a chance occurrence, 1,000 groups of 165 genes selected randomly were submitted to DAVID [19,20]. Among these 1,000 tests, only one gave better results for the total number of pathways than the 165 genes of interest. For the pathways related to cancer, only one test gave better results than the 165 genes of interest. For the pathways involved in cell proliferation and cell motility, only four tests gave equal or better results than the 165 genes of interest (Figure 3). This shows that the probability to obtain the results with the 165 genes of interest by chance is less than 0.5%.

Expression profiles

The expression values of the 165 genes of interest in three samples of non-metastatic breast cancer cells (MCF-7 cells) and two samples of metastatic breast cancer cells (MDA-MB-231) were used to construct expression profiles. Because the number of probe sets varied from one to six for each gene, the 165 genes generated 354 expression profiles (additional file 2). A large portion of them provide interesting information and validate certain results. Indeed, they can directly show up- or downregulation at the transcript level of the genes of interest between cancer cells with or without a metastatic phenotype. In addition, the expression values of four genes known to be involved in cell response to hypoxia in three replicates of metastatic pancreatic cancer cells (L3.6pl cells) incubated under hypoxia and three replicates of cells incubated under normoxia were used to construct expression profiles. They show the impact of hypoxia on the transcript level of these genes.

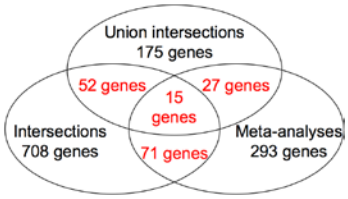


Figure 2: Venn's diagram of the selected genes. The intersections selected 846 genes, the union intersections selected 269 genes and the meta-analyses selected 406 genes. This data was used to generate a Venn's diagram.

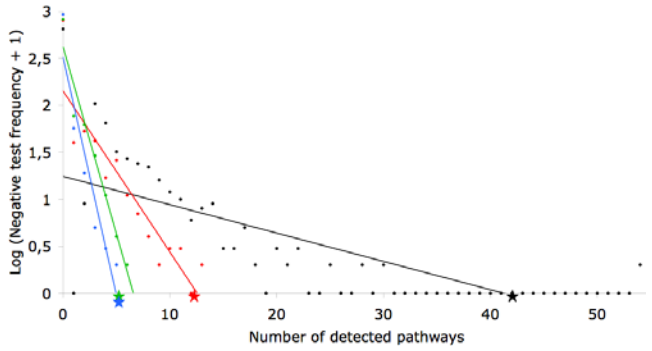


Figure 3: Pathways retrieved by DAVID in negative tests. 1,000 groups of 165 randomly-selected genes were submitted to DAVID. The X axis shows the number of pathways retrieved by DAVID and the Y axis shows the logarithm of the frequency of the tests (+ 1 to avoid log (0)). The total number of retrieved pathways is represented by the black dots for the negative tests and by the black star for the 165 genes of interest. The number of retrieved cancer pathways is represented by the red dots for the negative tests and by the red star for the 165 genes of interest. The number of retrieved pathways involved in proliferation and cell motility is represented by the green dots for the negative tests and by the green star for the 165 genes of interest. Finally, the number of retrieved pathways involved in pathogen recognition and phagocytosis is represented by the blue dots for the negative tests and by the blue star for the 165 genes of interest.

Discussion

The intersection approach is a promising method to perform a meta-analysis of a set of microarray datasets. It consists first of performing a classical analysis of all of the datasets, and then selecting the common genes to all the top gene lists of the datasets. This ensures that genes with high probability of differential expression are selected and that there are fewer false positives and false negatives among the genes selected. Moreover, this approach can be performed by re-exploiting archived datasets without performing new experiments. However, as far as we know and after consulting several biostatisticians, there is no statistic to calculate the probability that a gene would be selected in an intersection. Here, we have developed such a statistic. To explain it, we begin with a simple analogy [24]. We have 40 balls, among which 20 are red and numbered from 1 to 20, and we draw 10 balls. This situation is similar to a microarray experiment analysis since we select a small number of genes more or less differentially expressed (the red balls) from a larger number of genes, generally on the order of tens of thousands, that are not differentially expressed (the other balls). The probability to always draw a particular red ball in first position in the null hypothesis where a red ball has the same chance to be drawn as the other balls is equal to $\left(\frac{1}{40}\right)^{10}$. The probability to always draw a particular red ball in first or second position in the null hypothesis is equal to $\left(\frac{2}{40}\right)^{10}$. Hence, with N balls among which n are red, the probability p to always draw a particular red ball in at least r position (where $r \leq N$) in the null hypothesis in k draws is equal to equation (2):

$$p = \left(\frac{r}{N}\right)^{k} \quad (2)$$

	Pathways	Databases	Genes
Cancer	Prostate cancer	KEGG	CCNE2, FGFR1, HSP90AA1, IGF1, MAPK1, MAP2K1, NFKBIA, PIK3CD
	Pathways in cancer	KEGG	CTBP2, CCNE2, FGFR1, FZD1, HSP90AA1, IGF1, MAPK1, MAP2K1, NFKBIA, PIK3CD, STAT1, RALA
	Melanoma	KEGG	FGFR1, IGF1, MAPK1, MAP2K1, PIK3CD
	Pancreatic cancer	KEGG	MAPK1, MAP2K1, PIK3CD, STAT1, RALA
	Chronic myeloid leukemia	KEGG	CTBP2, MAPK1, MAP2K1, NFKBIA, PIK3CD
	Glioma	KEGG	IGF1, MAPK1, MAP2K1, PIK3CD
	Colorectal cancer	KEGG	FZD1, MAPK1, MAP2K1, PIK3CD
	Endometrial cancer	KEGG	MAPK1, MAP2K1, PIK3CD
	Non-small cell lung cancer	KEGG	MAPK1, MAP2K1, PIK3CD
	Acute myeloid leukemia	KEGG	MAPK1, MAP2K1, PIK3CD
	Renal cell carcinoma	KEGG	MAPK1, MAP2K1, PIK3CD
	Small cell lung cancer	KEGG	CCNE2, NFKBIA, PIK3CD
Proliferation and cell motility	Focal adhesion	KEGG	CAV1, FLNC, IGF1, MAPK1, MAP2K1, PIK3CD, SPP1
	VEGF signaling pathway	KEGG	HSPB1, MAPK1, MAP2K1, PIK3CD
	MAPK signaling pathway	KEGG, BIOCARTA	DUSP4, FGFR1, FLNC, HSPB1, MAPK1, MAP2K1, NR4A1, NFKBIA, STAT1
	ErbB signaling pathway	KEGG	MAPK1, MAP2K1, PIK3CD
	Regulation of actin cytoskeleton	KEGG	FGFR1, MAPK1, MAP2K1, PIK3CD
Pathogen recognition and phagocytosis	Toll-like receptor signaling pathway	KEGG	MAPK1, MAP2K1, NFKBIA, PIK3CD, SPP1, STAT1
	fMLP induced chemokine gene expression in HMC-1 cells	BIOCARTA	MAPK1, MAP2K1, NFKBIA
	T Cell Receptor Signaling Pathway	KEGG	MAPK1, MAP2K1, NFKBIA, PIK3CD
	Fc Epsilon Receptor I Signaling pathway	KEGG	MAPK1, MAP2K1, PIK3CD
	Fc gamma R-mediated phagocytosis	KEGG	MAPK1, MAP2K1, PIK3CD
Other	p53 signaling pathway	KEGG	CCNB1, CCNE2, IGF1, RRM2, SERPINE1, SFN
	Oocyte meiosis	KEGG	CCNB1, CCNE2, IGF1, MAPK1, MAP2K1, YWHAB, YWHAZ
	Cell cycle	KEGG	CDC6, CCNB1, CCNE2, SFN, YWHAB, YWHAZ
	Glutathione metabolism	KEGG	GGT1, GCLM, GSTM1, GSTM2, GSTM4, RRM2
	Metabolism of xenobiotics by cytochrome P450	KEGG	GSTM1, GSTM2, GSTM4
	Drug metabolism	KEGG	GSTM1, GSTM2, GSTM4
	Progesterone-mediated oocyte maturation	KEGG	CCNB1, HSP90AA1, IGF1, MAPK1, MAP2K1, PIK3CD
	Aldosterone-regulated sodium reabsorption	KEGG	IGF1, MAPK1, PIK3CD, SFN
	Neurotrophin signaling pathway	KEGG	MAPK1, MAP2K1, NFKBIA, PIK3CD, YWHAB, YWHAZ
	Cadmium induces DNA synthesis and proliferation in macrophages	BIOCARTA	MAPK1, MAP2K1, NFKBIA
	mTOR signaling pathway	KEGG	DDIT4, IGF1, MAPK1, PIK3CD
	NOD-like receptor signaling pathway	KEGG	CXCL1, HSP90AA1, MAPK1, NFKBIA
	B cell receptor signaling pathway	KEGG	MAPK1, MAP2K1, NFKBIA, PIK3CD
	Chemokine signaling pathway	KEGG	CXCL1, MAPK1, MAP2K1, NFKBIA, PIK3CD, STAT1
	NFAT and Hypertrophy of the heart (Transcription in the broken heart)	BIOCARTA	IGF1, MAPK1, MAP2K1
	Keratinocyte Differentiation	BIOCARTA	MAPK1, MAP2K1, NFKBIA
	Long-term depression	KEGG	IGF1, MAPK1, MAP2K1
	Natural killer cell mediated cytotoxicity	KEGG	HLA-C, MAPK1, MAP2K1, PIK3CD
	Melanogenesis	KEGG	FZD1, MAPK1, MAP2K1
	Insulin signaling pathway	KEGG	MAPK1, MAP2K1, PIK3CD

The 165 genes of interest were classified by DAVID into 42 different KEGG or Biocarta pathway maps

Table 1: Pathways retrieved by DAVID.

Hence, the probability p to not draw this ball in at least r position in the null hypothesis is equal to equation (3):

$$p = 1 - \left(\frac{r}{N}\right)^k \quad (3)$$

Hence, the probability p to not draw any red ball in at least r position in the null hypothesis is equal to equation (4):

$$p = \left[1 - \left(\frac{r}{N}\right)^k\right]^n \quad (4)$$

In conclusion, the probability p to draw at least one red ball in at least r position in the null hypothesis is equal to equation (5):

$$p = 1 - \left[1 - \left(\frac{r}{N}\right)^k\right]^n \quad (5)$$

Applied to a meta-analysis by the intersection approach of a set of microarray datasets, p represents the probability that a gene will be selected, r the maximum rank of a gene to be selected with the probability p, N the number of genes represented on the microarray, k the number of datasets taken into account in the intersection and n the number of genes likely to be involved in the phenomenon studied. Equation (5) thus calculates a maximum rank to select genes by the intersection approach with a chosen probability among several datasets (equation 6).

$$r = [1 - (1 - p)^{1/n}]^{1/k} \times N \quad (6)$$

In the meta-analysis presented in this paper, all the datasets were generated with Affymetrix platforms and, since some intersections included several GeneChip models, N in equation (6) was defined as the number of probe sets in the GeneChip model with the largest number of probe sets. This ensured calculation of the probability p. Indeed, if

N was not defined as the number of probe sets of the GeneChip model with the largest number of probe sets, this could lead to a rank r larger than the number of probe sets in some GeneChips of the intersection. We defined n as the number of genes likely to be involved in metastasis and/or in the response to hypoxia, whether or not described previously in the literature. Indeed, given the lack of evidence of the involvement of some genes in these phenomena, we were forced to estimate their number. For this, we consulted the Entrez Gene database (NCBI) to determine the number of genes already listed to be involved in metastasis and/or in the response to hypoxia. There are 710 genes responding to the keyword "metastasis", 480 responding to the keyword "hypoxia" and 134 responding to the keywords "metastasis AND hypoxia". Hence, we considered that there are 1,056 (710 + 480 – 134) genes known in the literature to be involved in metastasis and/or hypoxia. To obtain n, we multiplied this number by two to take into account the genes involved in metastasis and/or hypoxia still not demonstrated as such. The choice of multiplier was motivated by our observation in a previous work about cancer using a meta-analysis methodology in which we retrieved 183 genes of interest, among which 99 were already known to be involved in cancer [15]. This showed that the number of genes known to be involved in a cancer-related phenomenon can be multiplied by two to take into account those genes still not known to be involved. According to this reasoning, n was set to 2,112 (1,056 x 2) and p was set to 0.05.

The intersection approach retrieved 846 different genes. Since there are too many genes to process, we then added two supplementary approaches: union intersections and meta-analyses. In the end, 165 genes of interest were selected by the combination of the three approaches (Figure 3). Among these 165 genes, 41 were already known to be involved in the metastatic phenotype and 20 in the response to hypoxia (additional file 1). Here, we present detailed information about the up- or downregulation in metastasis or hypoxia and consistent expression profiles for 8 of the genes known to be involved in metastasis and 4 of the genes known to be involved in the response to hypoxia.

These genes include NR4A1 (nuclear receptor subfamily 4, group A, member 1) which is a nuclear receptor involved in cell differentiation,

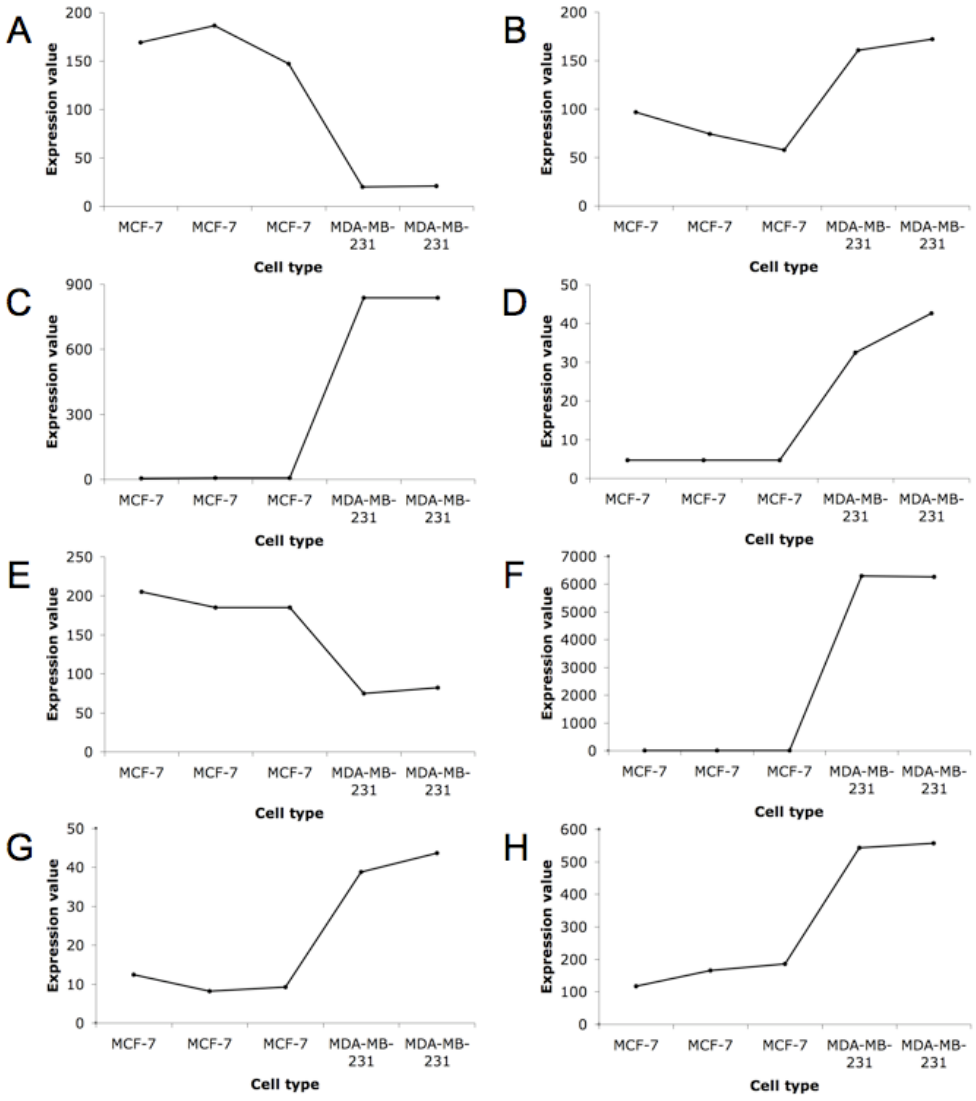


Figure 4: Expression profiles of genes involved in metastasis. The X axis shows the two cell types compared: MCF-7 and MDA-MB-231. The Y axis shows the expression value that reflects the transcript level of the (A) NR4A1, (B) ASPM, (C) BCAT1, (D) RGS20, (E) ZFHX3, (F) CAV1, (G) GPC1 and (H) SMURF2 genes in the sample.

proliferation and survival. Moreover, by migrating to mitochondria, NR4A1 allows BCL-2 (B-cell CLL/lymphoma 2) to trigger apoptosis. This is the reason why downregulation of NR4A1 leads to metastasis of cancer cells as they escape from apoptosis [25]. NR4A1 is clearly less expressed in MDA-MB-231 cells as shown in Figure 4A.

ASPM (asp (abnormal spindle) homolog, microcephaly associated (Drosophila)) is another gene retrieved by the methodology. Mutations of ASPM are responsible for microcephaly. Recent studies have suggested ASPM as an actor in the cell cycle and cell proliferation [26]. However, the diversity of its domains also suggests a large variety of biological functions. Overexpression of ASPM has been demonstrated to be a marker of metastasis as confirmed by the expression profile shown in Figure 4B.

BCAT1 (branched chain amino-acid transaminase 1, cytosolic) is also a gene of interest that shows clear upregulation in MDA-MB-231 cells in the expression profile in Figure 4C. BCAT1 codes for an enzyme responsible for the transamination of branched-chain alpha-keto acids to branched-chain L-amino acids occurring during cell growth. The upregulation of BCAT1 is a predictive factor for the development of metastases [27].

RGS20 (regulator of G-protein signaling 20) is a GTPase-activating protein involved in the regulation of signal transduction. A recent study showed a higher level of transcripts in metastatic melanomas than in primary melanomas [28]. The same observation was made between metastatic breast cancer cells and non-metastatic breast cancer cells in the expression profile in Figure 4D.

ZFHX3 (zinc finger homeobox 3) is another example of a gene selected by the methodology and already known to be involved in metastasis. Indeed, the protein encoded by ZFHX3 is a transcription factor that mediates cell differentiation and growth. It appears that ZFHX3 inhibits AFP (alpha-fetoprotein), which is often over-expressed in extremely malignant gastric cancers [29]. Thus, it is not surprising that transcript levels of ZFHX3 are very low in metastatic cancer cells as shown in the expression profile in Figure 4E.

CAV1 (caveolin 1, caveolae protein, 22kDa) plays a role in the formation of caveolae that are small lipid rafts responsible for vesicle trafficking, cholesterol homeostasis and signal transduction. High CAV1 levels have been linked with the metastatic phenotype as they lead to the secretion of MMP3 and MMP11 (matrix metalloproteinases 3 and 11) [30]. High CAV1 levels are also observed in MDA-MB-231 cells compared to MCF-7 cells in the expression profile in Figure 4F.

GPC1 (glypican 1) is a HBGF (heparin-binding growth factor) coreceptor found by the methodology and that shows high upregulation in metastatic breast cancer cells in the expression profile in Figure 4G. Studies have demonstrated that high levels of GPC1 lead to cancer metastasis [31].

As a last example of a gene selected by the methodology and validated by an expression profile (Figure 4H) comparing metastatic and non metastatic cancer cells, SMURF2 (SMAD specific E3 ubiquitin protein ligase 2) is an E3 ligase that induces a modification of ubiquitin to thus modulate the TNF-β (tumor necrosis factor) signal. High transcript levels of SMURF2 have been shown to be associated with high metastatic potential [32].

ADM (adrenomedullin) is the first example of a gene known to be involved in the response to hypoxia that was selected by the methodology. Figure 5A presents the upregulation of ADM under hypoxia. The ADM protein has been found in many cell types in different tissues such as the heart, lung, kidney and pancreas. Studies have shown that this protein has several functions including proliferation, differentiation, migration and regulation of blood pressure. Observations of upregulation under hypoxia, anti-apoptotic effects and promotion of angiogenesis suggest that ADM could be a major actor in the development of cancer [33].

MAP2K1 (mitogen-activated protein kinase kinase 1) is a kinase of the MAPK signal transduction pathway that is involved in various signaling of extracellular signals inside the cell. MAP2K1 can be activated through hypoxia to trigger cellular processes such as proliferation, migration and survival [34]. Figure 5B shows the upregulation of MAP2K1 under hypoxia.

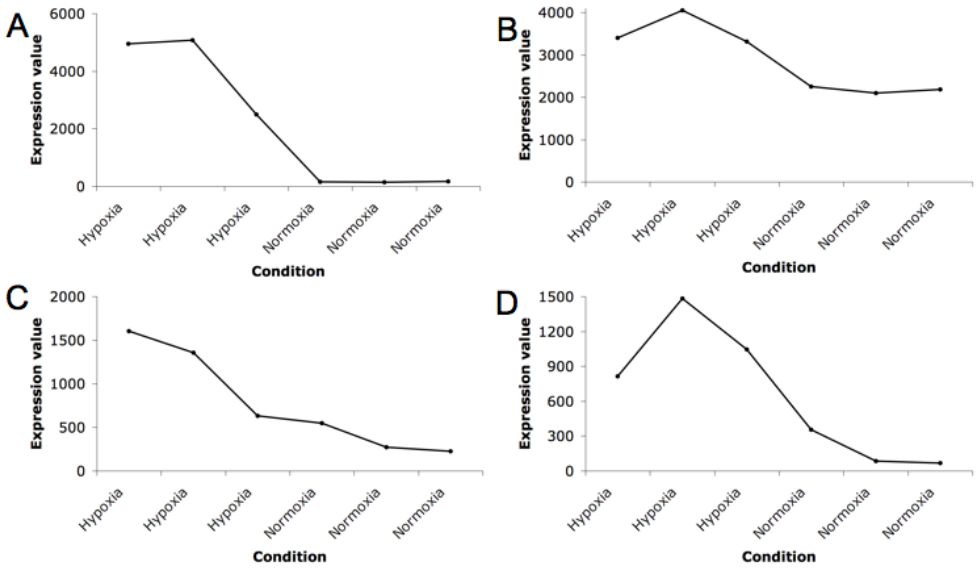


Figure 5: Expression profiles of genes involved in hypoxia. The X axis shows the two conditions compared: hypoxia and normoxia in metastatic pancreatic cancer cells (L3.6pl). The Y axis shows the expression value that reflects the transcript level of the (A) ADM, (B) MAP2K1, (C) BHLHE40 and (D) SLC2A3 genes in the sample.

The transcription repressor BHLHE40 (basic helix-loop-helix family, member e40), which is a member of the bHLH leucine zipper family, shows upregulation under hypoxic condition in the expression profile (Figure 5C). This is consistent with the fact that BHLHE40 is a target of the HIF protein family and that in a lack of oxygen, overall transcription is repressed to save energy. The dimer form of BHLHE40 is involved in cell differentiation, circadian rhythms, immune regulation and carcinogenesis [35].

In conclusion, as a last example of a gene selected by the methodology and validated by an expression profile comparing hypoxia and normoxia in pancreatic cancer cells, we describe SLC2A3 (solute carrier family 2, member 3), better known as GLUT3 (glucose transporter type 3), a glucose carrier system. Surprinsingly, unlike GLUT1, GLUT3 is specific to neurons, but was picked out by the methodology. Studies have shown that the lack of oxygen or glucose is responsible for the upregulation of GLUT3, as presented in Figure 5D [36]. This upregulation allows the cell to switch from aerobic to anaerobic metabolism.

Many of the genes selected by the methodology are thus known to be involved in cancer, metastasis and/or hypoxia (additional file 1). These genes were classified into KEGG [21] and Biocarta (Biocarta) pathways by DAVID [19,20] and the same observation made at the gene level can also be made at the pathway level as many of these pathways are related to cancer or cell proliferation and motility. Indeed, DAVID retrieved 42 different pathways from the 165 genes of interest (Table 1) and a negative control composed of 1,000 tests demonstrated that there is less than a 1‰ chance of obtaining such results by chance (Figure 3). Among the 42 pathways, 12 are cancer pathways (Table 1). These 12 cancer pathways are "prostate cancer", "pathways in cancer", "melanoma", "pancreatic cancer", "chronic myeloid leukemia", "glioma", "colorectal cancer", "endometrial cancer", "non-small cell lung cancer", "acute myeloid leukemia", "renal cell carcinoma" and "small cell lung cancer". All of these pathways are directly related to cancer since they reflect the molecular interactions of some types of cancer. The number of genes selected by the methodology and involved in these pathways varies between 3 and 12.

Taken together, these various arguments indicate the power of the proposed methodology. Hence, we suggest that the 74 genes (165 – 91) still not described to be implicated in cancer are potential new factors of tumor growth and particularly of metastasis induced by hypoxia. Surprisingly, DAVID [19,20] retrieved five pathways from the 165 genes of interest which are related to pathogen recognition and phagocytosis (Table 1). These pathways are "toll-like receptor signaling pathway", "fMLP induced chemokine gene expression in HMC-1 cells", "T cell receptor signaling pathway", "Fc epsilon RI signaling pathway" and "Fc gamma R-mediated phagocytosis". The first four were already discussed previously in Pierre et al. [15]. However, the Fc gamma R-mediated phagocytosis pathway was not. Fc gamma R-mediated phagocytosis is a major process set up by macrophages, neutrophils and monocytes to eliminate a pathogen threat. Following the extracellular recognition of a pathogen molecule by an Fc gamma receptor, an intracellular signal induces the development of a phagosome that then merges with lysosomes. Lysosomal proteases digest the pathogen. It is interesting to note that development of the phagosome requires regulation of the actin cytoskeleton [37,38], which was a pathway also retrieved by DAVID from the 165 genes of interest. This involvement could be the link between metastasis and Fc gamma R-mediated phagocytosis.

Though these pathways have not been previously reported to be involved in cancer, metastasis or hypoxia, they were selected by the methodology, and the 1,000 negative tests demonstrate that the probability to obtain five pathways related to pathogen recognition and phagocytosis is equal to 1‰ since only one test gave results equal to those with the 165 genes of interest (Figure 3).

Further steps of this work include *in vitro* validation of the expression of the genes of interest in MDA-MB-231 and MCF-7 cell lines and functional analyses of the proteins encoded by the genes of interest. These approaches should open new doors to understand the metastasis process under hypoxic conditions.

We propose in this paper a major advance in a meta-analysis methodology. Here, we report the development and application of a statistic that sets a statistical threshold to the proposed approach, hence eliminating the need to make an arbitrary choice. In addition to our observation of results consistent with the studied phenomenon, a large negative control consisting of 1,000 random tests and two independent validations of expression profiles of the genes of interest support the ability of the methodology not only to retrieve genes already known to be involved in the phenomenon but to identify new reliable candidate genes.

Acknowledgements

M. Pierre is supported by FRIA (Belgium), B. DeMeulder is supported by Televie (Belgium) and S. Depiereux is supported by the FNRS (Belgian National Scientific Research Fund). We would like to thank P. Dagnelie for helpful discussions in the development of the intersection statistic. We would also like to thank J.J. LaPres (Biochemistry and Molecular Biology, Michigan State University, East Lansing) for providing the dataset GSE1056 and K.S. Hoek (Department of Dermatology, University Hospital of Zürich, Zürich) for providing the datasets GSE4840 and GSE4843.

References

1. Friedl P, Wolf K (2003) Tumour-cell invasion and migration: diversity and escape mechanisms. Nat Rev Cancer 3: 362-374.
2. Pantel K, Brakenhoff RH (2004) Dissecting the metastatic cascade. Nat Rev Cancer 4: 448-456.
3. Chan DA, Giaccia AJ (2007) Hypoxia, gene expression, and metastasis. Cancer Metastasis Rev 26: 333-339.
4. Sullivan R, Graham CH (2007) Hypoxia-driven selection of the metastatic phenotype. Cancer Metastasis Rev 26: 319-331.
5. Vaupel P (2004) The role of hypoxia-induced factors in tumor progression. Oncologist 9 Suppl 5: 10-17.
6. DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278: 680-686.
7. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society 57: 289-300.
8. Baldi P, Long AD (2001) A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics 17: 509-519.
9. Gautier L, Moller M, Friis-Hansen L, Knudsen S (2004) Alternative mapping of probes to genes for Affymetrix chips. BMC Bioinformatics 5: 111.
10. Liu H, Zeeberg BR, Qu G, Koru AG, Ferrucci A, et al. (2007) AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets. Bioinformatics 23: 2385-2390.
11. Gur-Dedeoglu B, Konu O, Kir S, Ozturk AR, Bozkurt B, et al. (2008) A resampling-based meta-analysis for detection of differential gene expression in breast cancer. BMC Cancer 8: 396.
12. Ma S, Huang J (2009) Regularized gene selection in cancer microarray meta-analysis. BMC Bioinformatics 10: 1.
13. Ochsner SA, Steffen DL, Hilsenbeck SG, Chen ES, Watkins C, et al. (2009) GEMS (Gene Expression MetaSignatures), a Web resource for querying meta-analysis of expression microarray datasets: 17beta-estradiol in MCF-7 cells. Cancer Res 69: 23-26.

14. Hunter KW, Crawford NP, Alsarraj J (2008) Mechanisms of metastasis. Breast Cancer Res 10 Suppl 1: S2.

15. Pierre M, DeHertogh B, Gaigneaux A, DeMeulder B, Berger F, et al. (2010) Meta-analysis of archived DNA microarrays identifies genes regulated by hypoxia and involved in a metastatic phenotype in cancer cells. BMC Cancer 10: 176.

16. Wu Z, Irizarry R, Gentleman R, Murillo F, Spencer F (2004) A model-based background adjustment for oligonucleotide expression arrays. Journal of the American Statistical Association 99: 909-917.

17. Berger F, De Hertogh B, Pierre M, Gaigneaux A, Depiereux E (2008) The "Window t test": a simple and powerful approach to detect differentially expressed genes in microarray datasets. Central European Journal of Biology 3: 327-344.

18. De Hertogh B, De Meulder B, Berger F, Pierre M, Bareke E, et al. (2010) A benchmark for statistical microarray data analysis that preserves actual biological and technical variance. BMC Bioinformatics 11: 17.

19. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4: P3.

20. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.

21. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27: 29-34.

22. Ihaka R, Gentleman R (1996) R : a language for data analysis and graphics. Journal of Computational and Graphical Statistics 5: 299-314.

23. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.

24. Dagnelie P (2007) Statistique théorique et appliquée. Bruxelles: De Boeck et Larcier.

25. Liu J, Zhou W, Li SS, Sun Z, Lin B, et al. (2008) Modulation of orphan nuclear receptor Nur77-mediated apoptotic pathway by acetylshikonin and analogues. Cancer Res 68: 8871-8880.

26. Lin SY, Pan HW, Liu SH, Jeng YM, Hu FC, et al. (2008) ASPM is a novel marker for vascular invasion, early recurrence, and poor prognosis of hepatocellular carcinoma. Clin Cancer Res 14: 4814-4820.

27. Yoshikawa R, Yanagi H, Shen CS, Fujiwara Y, Noda M, et al. (2006) ECA39 is a novel distant metastasis-related biomarker in colorectal cancer. World J Gastroenterol 12: 5884-5889.

28. Riker AI, Enkemann SA, Fodstad O, Liu S, Ren S, et al. (2008) The gene expression profiles of primary and metastatic melanoma yields a transition point of tumor progression and metastasis. BMC Med Genomics 1: 13.

29. Zhang Z, Yamashita H, Toyama T, Sugiura H, Ando Y, et al. (2005) ATBF1-a messenger RNA expression is correlated with better prognosis in breast cancer. Clin Cancer Res 11: 193-198.

30. Du ZM, Hu CF, Shao Q, Huang MY, Kou CW, et al. (2009) Upregulation of caveolin-1 and CD147 expression in nasopharyngeal carcinoma enhanced tumor cell migration and correlated with poor prognosis of the patients. Int J Cancer 125: 1832-1841.

31. Aikawa T, Whipple CA, Lopez ME, Gunn J, Young A, et al. (2008) Glypican-1 modulates the angiogenic and metastatic potential of human and mouse cancer cells. J Clin Invest 118: 89-99.

32. Jin C, Yang YA, Anver MR, Morris N, Wang X, et al. (2009) Smad ubiquitination regulatory factor 2 promotes metastasis of breast cancer cells by enhancing migration and invasiveness. Cancer Res 69: 735-740.

33. Keleg S, Kayed H, Jiang X, Penzel R, Giese T, et al. (2007) Adrenomedullin is induced by hypoxia and enhances pancreatic cancer cell invasion. Int J Cancer 121: 21-32.

34. Wang K, Jiang YZ, Chen DB, Zheng J (2009) Hypoxia enhances FGF2- and VEGF-stimulated human placental artery endothelial cell proliferation: roles of MEK1/2/ERK1/2 and PI3K/AKT1 pathways. Placenta 30: 1045-1051.

35. Choi SM, Cho HJ, Cho H, Kim KH, Kim JB, et al. (2008) Stra13/DEC1 and DEC2 inhibit sterol regulatory element binding protein-1c in a hypoxia-inducible factor-dependent mechanism. Nucleic Acids Res 36: 6372-6385.

36. Bruckner BA, Ammini CV, Otal MP, Raizada MK, Stacpoole PW (1999) Regulation of brain glucose transporters by glucose and oxygen deprivation. Metabolism 48: 422-431.

37. May RC, Machesky LM (2001) Phagocytosis and the actin cytoskeleton. J Cell Sci 114: 1061-1077.

38. Groves E, Dart AE, Covarelli V, Caron E (2008) Molecular mechanisms of phagocytic uptake in mammalian cells. Cell Mol Life Sci 65: 1957-1976.

39. Biocarta. Biocarta Pathways.

40. NCBI. Entrez Global Query Cross-Database Search System.

41. NCBI. 2000. Gene Expression Omnibus.

2. Résultats *in vitro*

Les résultats suivants sont présentés sous forme d'un article à soumettre. Le format qui a été choisi est celui du groupe BMC, mais ce format pourrait être amené à changer en fonction de la revue qui sera finalement choisie pour soumettre l'article. La présente forme regroupe toutes les validations *in vitro* qui ont été réalisées pour confirmer les hypothèses générées par la méthodologie bioinformatique décrite plus haut.

Il s'agit de la validation par RT-PCR en temps réel, dans des cellules à haut et à bas potentiel métastatique, du niveau d'expression de neuf gènes impliqués dans la reconnaissance de pathogène et la phagocytose. De plus, des tests de migration dans des cellules à haut potentiel métastatique, dans lesquelles l'expression de l'un de deux de ces neuf gènes (PAK1 et CFL2) a été invalidée par siRNA, ont montré une implication dans la migration de ces cellules cancéreuses.

Enfin, un test de viabilité cellulaire a montré que la perte de capacité migratoire suite à l'invalidation de PAK1 ou de CFL2 n'était pas dû à une chute de la prolifération de ces cellules. PAK1 et CFL2 représentent dès lors des candidats intéressants pour de futures thérapies contre le développement métastatique.

PAK1 and CFL2 promote the metastatic phenotype of cancer cells through the Fc Gamma R-mediated phagocytosis pathway

Michael Pierre[1], Annick Notte[2], Lionel Leclere[2], Kayleen Vannuvel[2], Eric Depiereux[1] and Carine Michiels[2§]

[1]Molecular Biology Research Unit (URBM), University of Namur - FUNDP, Namur, Belgium

[2]Cell Biology Research Unit (URBC), NARILIS, University of Namur - FUNDP, Namur, Belgium

[§]Corresponding author: 61 rue de Bruxelles, 5000 Namur, Belgium, tel.: +32 81 72 41 31, fax: +32 81 72 41 35

Email addresses:

MP: michael.pierre@fundp.ac.be

AN: anotte@fundp.ac.be

LL: lionel.leclere@fundp.ac.be

KV: kayleen.vannuvel@fundp.ac.be

ED: eric.depiereux@fundp.ac.be

CM: carine.michiels@fundp.ac.be

Abstract

Background

Metastasis characterizes the final stage of cancer. Cancer cells are then able to migrate to distant organs to produce secondary tumors. This phenomenon requires the setup of several molecular mechanisms inside and outside the cancer cells. Along with these mechanisms, the expression profile of many genes changes. Studies demonstrated that metastasis is enhanced by hypoxia that develops at the center of the tumor. In a previous study, we picked out, through a DNA microarray meta-analysis, genes that are down- or upregulated by hypoxia and potentially involved in metastasis. In this paper, we present *in vitro* validations of these hypotheses generated *in silico*.

Methods

Expression profiles have been generated from a DNA microarray dataset comparing MDA-MB-231 cells with a high metastatic potential and MCF-7 cells with a low metastatic potential for the genes of the Fc Gamma R-mediated phagocytosis pathway and the toll like receptor signalling pathway. The transcript level of 9 genes (VAV2, RAC2, PAK1, LIMK1, CFL2, TICAM1, TRAF3, TBK1 and IRF3) has been assayed by real-time RT-PCR in MDA-MB-231 and MCF-7 cells. MDA-MB-231 cells have been transfected with siRNAs targeting PAK1 or CFL2 and invalidation assessed by real-time RT-PCRs and western blot. Migration and viability were respectively evaluated by scratch and MTT assays.

Results

In silico expression profiles of the genes of the Fc Gamma R-mediated phagocytosis pathway and the toll like receptor signalling pathway indicate a possible role for both pathways in metastasis. Moreover, real-time RT-PCR assays performed on MCF-7 compared with MDA-MB-231 cells validate this observation for four genes from the two pathways. MDA-MB-231 cells where PAK1 or CFL2 expression was invalidated showed an important decrease in their ability to migrate as early as 48 h post-transfection.

Conclusion

The results presented here indicate that PAK1 and CFL2 are upregulated during the metastatic process and that their invalidation in highly metastatic cancer cells decreases their motility. PAK1 and CFL2 are therefore interesting new targets for cancer therapy.

Background

In the final stage of cancer, tumor cells display the ability to migrate to organs distant from the primary tumor [1]. This phenomenon is called metastasis. To trigger the metastasis capability, several molecular mechanisms take place inside and outside the cancer cells [2]. These mechanisms allow these cells to detach from the other cancer cells and to adhere to the extracellular matrix, to then degrade it. Next, the migrant cells have to penetrate inside a blood or a lymph vessel to reach a body area where space and nutrients are in sufficient amounts. In this area, they have to go out from the blood or lymph vessels and divide in the distant organ. Changes in the expression profile of specific genes, such as integrins, matrix metalloproteinases and growth factors, are observed during the different steps of metastasis.

Hypoxia and metastasis are closely related. Indeed, as the tumor is growing, cancer cells become more and more distant from the blood vessels. This phenomenon results in a decrease in oxygen supply. Even if the cancer cells induce the development of a new vasculature inside the tumor, this one is abnormal and delivers only few or irregular amounts of oxygen [3]. Studies showed that hypoxia activates the transcription factor HIF-1 resulting in an increased cancer cell survival [4] and in the angiogenesis process. Moreover, hypoxia is known to decrease the effect of radiotherapy and chemotherapy [5]. Finally, it seems that highly metastatic cancer cells are selected by hypoxia [6]. All these evidences indicate that hypoxia is a marker for poor prognosis.

In two previous works [7, 8], we proposed a methodology to achieve a meta-analysis of DNA microarray data related to metastasis and hypoxia. DNA microarrays allow to assess the whole transcriptome of cells or biological samples and their comparison from a condition to another [9]. However, several issues make the analysis of the data they generate tricky. These issues include the large number of false positives and false negatives generated by the technique [10], the lack of correspondence between the probes and the genes they are expected to detect [11] and the too large choice of methods to achieve quantitative analysis of the data. However, the increasing number of publicly available datasets led to a new way to analyze them: it is the meta-analysis. This involves analyzing several datasets at once to increase the statistical power of the analysis [12-14].

Our meta-analysis methodology allowed us to highlight genes and pathways that were not previously described as involved in metastasis. In this paper, we investigate these hypotheses generated *in silico* using *in vitro* experiments. In addition to validate the meta-analysis

methodology, the results presented here indicate an important role of two genes (PAK1 and CFL2) in the metastatic phenotype of cancer cells.

Methods

Expression profiles

The DNA microarray dataset GSE5823 was downloaded from GEO [15]. It was analyzed with a transcript-consistent AffyProbeMiner's CDF [16] and pre-processed with GCRMA [17] with default parameters. The expression values of each probe set for each gene of interest were then plotted for GeneChips where control MCF-7 cells or control MDA-MB-231 cells were analyzed.

Cell cultures

Human breast cancer MDA-MB-231 and MCF-7 cells were maintained separately in Roswell Park Memorial Institute (RPMI) 1640 medium (Gibco, Paisley, UK) containing 10% fetal calf serum at 37°C and 5% CO_2.

Total RNA extraction

MDA-MB-231 and MCF-7 cells were grown separately in T75 flasks. The medium was removed and total RNA extraction was performed using RNAgents kit according to the manufacturer's instructions (RNAgents, Total RNA Isolation System, Promega, Madison). Total RNA was then stored at -70°C.

Reverse transcription

For each cell type, 1 µg of total RNA was diluted in 12 µl of water. 1 µl of Anchored-oligo (dT) 18 Primer (50 pmol/µL) (Roche) was then added. This mix was incubated for 10 min at 65°C. 7 µl of reaction mix [4 µl of Transcriptor Reverse Transcriptase Reaction Buffer 5X (Roche); 0.5 µl of Protector RNase Inhibitor (40 U/µl) (Roche); 2 µl of Deoxynucleotide Mix 10 mM (Roche); 0.5 µl of Transcriptor Reverse Transcriptase (20 U/µl) (Roche)] were added and the samples were incubated 30 min at 55°C, then 5 min at 85°C and finally 5 min on ice before to be stored at -20°C.

Real-time RT-PCR

Sequences of the primers were determined using Primer Express 1.5 software (Applied Biosystems, Foster City): TICAM1, 5'-TGCACAGGCCCATCACTTC-3' (forward) and 5'-AGTTTGTGCTTCAGATACAAGAGCTT-3' (reverse); TRAF3, 5'-TCGAAATAATGAATCCAAAATCCTT-3' (forward) and 5'-TCTCCTTGTCAAGCTCCTTCAGT-3' (reverse); TBK1, 5'-CGCACTTTACAGATGAATGTGTTAAA-3' (forward) and 5'-GCGATAATAACTGTTTCCTAAGATGAAG-3' (reverse); IRF3, 5'-AAGGAAGGAGGCGTGTTTGA-3' (forward) and 5'-CCTTCCGTGAAGGTAATCAGATCT-3' (reverse); VAV2, 5'-ACCACACTCAAGTACCCCTACAAGT-3' (forward) and 5'-GGACTGAGAAAAGAAAAGTTGTAGGAA-3' (reverse); RAC2, 5'-AAGCTGGCTCCCATCACCTA-3' (forward) and 5'-TGAGAGCTGAGCACTCCAGGTA-3' (reverse); PAK1, 5'-GGATGAAGGCCAAATTGCA-3' (forward) and 5'-TGTGAATGACCTGGTTCGAATG-3' (reverse); LIMK1, 5'-ATCATCCACCGAGACCTCAACT-3' (forward) and 5'-GTCAGCCACCACCACATTCTT-3' (reverse); CFL2, 5'-CATACGAAACAAAAGAGTCTAAGAAAGAA-3' (forward) and 5'-CATCTTTAGAGCTAGCATAAATCATCTTG-3' (reverse). 5 µl of 100X diluted cDNA previously obtained by reverse transcription of total RNA were mixed to SYBR Green Master Mix PCR [5 µl of distilled water; 0.84 µl of primer Reverse at 9 µM; 0.84 µl of primer Forward at 9 µM; 12.5 µl of SYBR green]. PCRs were carried out in a real-time PCR cycler (ABI PRISM 7700 Sequence Detector, PE Applied Biosystems). Thermal cycling conditions were: initial incubation of 10 min at 95°C, followed by 40 cycles of 30 s at 95°C, 1 min at 57°C annealing temperature, and 30 s at 72°C. Samples were compared using the relative Ct method [18]. To normalize for input load of cDNA between samples, the mean amplification of 23kDa and α-tubulin was used as the endogenous standard.

siRNA transfection

500,000 MDA-MB-231 cells were grown for 24 h in T25 flasks. Transfection medium was prepared: 10 µl of siRNA (20 µM) (ON-TARGET plus SMART pool, Thermo Scientific) were added to 390 µl of Opti-MEM medium, 8 µl of DharmaFECT were added to 392 µl of Opti-MEM medium. After 5 min, these two solutions were mixed together, and after 20 min, 3.2 ml of RPMI 1640 medium (Gibco, Paisley, UK) containing 10% fetal calf serum were

added. Cell medium was replaced by transfection medium in the T25 flasks. After 24 h, transfection medium was replaced by RPMI 1640 medium (Gibco, Paisley, UK) containing 10% fetal calf serum. Finally, after 24 h, cells were used for the appropriate experiments.

Scratch assays

24 h after the end of the transfection, small scratches were made in the cell monolayer using a scraper and cells were rinsed twice with PBS. The cell medium was then replaced by fresh medium. Pictures of the scratches were taken at 0 h, 6 h and 12 h after they have been made with a microscope (Leitz) coupled to a camera (DC-100, Leica). Quantization of cell migration was performed counting the number of cells by arbitrary surface unit on randomly captured pictures of 3 independent scratches.

Protein extraction

Total protein extracts were prepared from MDA-MB-231 cells grown in T25 flasks. After elimination of the medium, cells were scrapped in 200 µl of lysis buffer (Tris 80 mM pH 7.5, KCl 300 mM, EDTA 2 mM, triton X100 1%) containing a protease inhibitor mixture (« Complete » from Roche Molecular Biochemicals, 1 tablet in 2 ml H_2O, added at a 1 : 25 dilution) and phosphatase inhibitors ($NaVO_3$ 25 mM, PNPP 250 mM, β-glycerophosphate 250 mM and NaF 125 mM, at a 1 : 25 dilution). Lysates were transferred into microtubes and centrifugated 5 min at 13,000 rpm at 4°C. Supernatants were collected and stored at -70°C.

Western blot

Total protein extracts were separeted on 10% sodium dodecyl sulfate-poly-acrylamide gel electrophoresis (SDS-PAGE) gels and then transferred to a polyvinylidene difluoride membrane (Amersham Biosciences). After blocking in 4 ml of phosphate saline buffer supplemented with 4 ml of LiCor for 1 hour at room temperature under shaking, the membrane was put in phosphate saline buffer containing 0.1% (v/v) Tween, the blot was probed with the following antibodies: anti-CFL2 antibodies (Santa Cruz; diluted 1:100), secondary antibody anti-goat (diluted 1:7,500), anti-PAK1 antibodies (Cell Signalling; diluted 1:1,000), and secondary antibody anti-rabbit (diluted 1:7,500). Proteins were visualized by scanning the membrane on an Odyssey Infrared Imaging System (LI-COR Biosciences) with both 700- and 800-nm channels.

Cell viability assay

MDA-MB-231 cells were seeded at a density of 10,000 per well in a 24 well plate in RPMI 1640 medium (Gibco, Paisley, UK) containing 10% fetal calf serum and let grow for 24 h, 48 h, 72 h or 96 h (with medium replacement every 24 h). After the incubation, cell viability was evaluated by 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide reduction assay (MTT) (Sigma, St. Louis, MO). 500 μl of MTT solution was added to each well for 2 h at 37°C under 5% CO_2, and then discarded before 1 h of incubation in lysis buffer [20% SDS (MP Biomedicals, Eschwege, Germany), 33.3% N,N-dimethyl-formamide (Merck, Darmstadt, Germany), pH 4.7] under shaking (70 rpm) at 37°C in the dark. The OD values were recorded at 570 nm (plate reader Ultramark Microplate Imaging System, Bio-Rad, Munchen, Germany).

Results

Expression profiles

Expression profiles were generated from a DNA microarray dataset (GSE5823 downloaded from GEO) comparing a highly metastatic breast cancer cell type (MDA-MB-231) and a poorly metastatic breast cancer cell type (MCF-7). These expression profiles were generated for all the probe sets corresponding to the genes of the Fc Gamma R-mediated phagocytosis pathway and the toll like receptor signalling pathway (Figures 1 and 2). These two pathways were retrieved from the KEGG database [19]. Indeed, they were shown to be possibly involved in the metastatic phenotype in previous studies [7, 8]. The expression profiles are available in the additional files 1 and 2. Only the expression profiles of 9 of these genes (VAV2, RAC2, PAK1, LIMK1, CFL2, TICAM1, TRAF3, TBK1 and IRF3) are presented here (Figures 3 and 4). For these genes, marked up- or downregulation of their expression level can be observed in the highly metastatic cancer cells.

mRNA expression

To confirm the results of the expression profiles generated from the DNA microarray dataset, we measured the transcript levels of these genes by the real-time RT-PCR technique. MDA-MB-231 and MCF-7 cells were cultured separately, their total RNA was extracted, then reverse transcribed. The cDNA of the genes of interest was measured and compared. To allow comparison of the expression between cell types, the mean amplification of two house

keeping genes (23kDa and α-tubulin) was used as endogenous standards. This avoids considering only a single house keeping gene which amplification could slightly change in one cell type compared with the other leading to false measures for the other genes. Figures 5 and 6 show the fold induction of the genes in the MDA-MB-231 cells compared with the MCF-7 cells. Consistent results with the expression profiles generated from the DNA microarray dataset can be observed for RAC2, PAK1 and CFL2 in the Fc Gamma R-mediated phagocytosis pathway, and for IRF3 in the toll like receptor signalling pathway.

Gene invalidation

Since four of the nine genes of interest showed a similar change in gene expression in highly metastatic cancer cells versus poorly metastatic cells in both the GEO dataset and in our experimental data, we chose two of these four genes to assess their role in the regulation of cell migration since they belong to the same pathway. The effect of their invalidation was studied in cell migration using scratch assays.

To check if the siRNA transfection was really effective, the transcript levels of PAK1 and CFL2 were measured by real-time RT-PCR in comparison to a non-targeting siRNA used as a negative control. The non-targeting siRNA had no effect on the transcript level of PAK1 and CFL2. PAK1 siRNA decreased by 65% the transcript level of PAK1 and had no effect on the transcript levels of CFL2. CFL2 siRNA strongly decreased the trancript level of CFL2 and slightly decreased the transcript level of PAK1 (Figure 7).

As a second control of the siRNA transfection efficiency, the protein levels of PAK1 and CFL2 were assessed by Western blot in MDA-MB-231 cell cultures. The protein levels were also assessed in control cell cultures or in cells transfected with the non-targeting siRNA. Results showed that the protein level of PAK1 was efficiently decreased when cells were transfected with the PAK1 siRNA without being affected when CFL2 siRNA was used (Figure 8). It must be noted that no effective antibody was found to detect the protein level of CFL2 (data not shown).

Effect of PAK1 and CFL2 invalidation on cell migration

Scratch assays have been performed to determine if PAK1 or CFL2 was involved in regulating cancer cell migration. PAK1 or CFL2 was invalidated in the highly metastatic MDA-MB-231 cells. Non-targeting siRNA was used as negative control. Scratches were made in cell monolayer. Cell migration was controlled 0, 6 and 12 h after the scratches have

been made. Figure 9 presents the colonization of the scratches for the different cultures. Figure 10 presents the quantification of the cells present in the scratch. In each case, a decrease in the migration can be observed when PAK1 or CFL2 was invalidated suggesting a role of these genes in the metastatic phenotype of the MDA-MB-231 cells. No effect of the non targeting siRNA was detected.

Effect of PAK1 and CFL2 invalidation on cell viability

To check if the decrease in cell migration of MDA-MB-231 cells, where PAK1 or CFL2 expression was invalidated, was not due to a decrease in proliferation, proliferation assays were performed. Cell density was estimated by the MTT method. A decrease in cell number compared to untransfected cells was observed when cells were transfected with PAK1 or CFL2 siRNA but also with the non-targeting siRNA, indicating that this decrease was probably due to the transfection *per se* (Figure 11). There was however no further decrease than the one induced by the non-targeting siRNA, suggesting that the invalidation of PAK1 or CFL2 does not affect cell proliferation specifically but rather affects cell mobility.

Discussion

In two previous papers [7, 8], we presented a DNA microarray meta-analysis methodology. 22 DNA microarray datasets allowed us to highlight genes and pathways potentially involved in hypoxia-dependent metastasis. Most of these genes and pathways were not previously known to be involved in this process. Indeed, among the 165 genes retrieved by the methodology, 41 have already been cited in papers related to metastasis. And among the 42 pathways retrieved by DAVID from these 165 genes, 5 were already known to be involved with more or less effect in proliferation and cell motility. Interestingly, 5 other pathways were related to pathogen recognition and phagocytosis. These pathways are « Fc epsilon R1 signalling pathway », « Fc Gamma R-mediated Phagocytosis », « fMLP induced chemokine gene expression in HMC-1 cells », « T cell Receptor Signalling Pathway » and « Toll Like Receptor Signalling Pathway ». 1,000 negative tests composed of genes selected randomly demonstrated that the probability to obtain such a result by chance was less than 1‰. Then, we confirmed the expression profiles of all the genes included in these 5 pathways with an independent DNA microarray dataset comparing MDA-MB-231 and MCF-7 cells, respectively cells with a high and a low metastatic ability. We then decided to investigate in more details the Fc Gamma R-mediated phagocytosis and the toll like receptor signalling

pathways. Indeed, both of them present a sub-pathway where all the genes display an important up- or downregulation (Figures 1 to 4). The Fc Gamma R-mediated phagocytosis is triggered by macrophages, neutrophils and monocytes when a pathogen molecule binds to their Fc Gamma receptors. This binding induces an intracellular signal resulting in the development of a phagosome. The phagosome then merges with lysosomes to digest the pathogen [20, 21]. The toll like receptor signalling pathway is also triggered when a pathogen molecule binds to a toll like receptor of an immune response cell. When it happens, inflammatory genes are induced resulting in an appropriate cellular response to eliminate the pathogen [22].

Since the expression profiles compared MDA-MB-231 and MCF-7 cells, we decided to experimentally validate these results for the two sub-pathways by real-time RT-PCR on the same cell types. The sub-pathway of the Fc Gamma R-mediated phagocytosis includes 5 genes : VAV2, RAC2, PAK1, LIMK1 and CFL2 (Figure 1). The sub-pathway of the toll like receptor signalling pathway includes 4 genes : TICAM1, TRAF3, TBK1 and IRF3 (Figure 2). To allow the comparison of the expression of genes through two cell types, we chose to use two house keeping genes : 23kDa and α-tubulin. We computed the mean value of their amplification and used it as the endogenous standard. This avoids relying only on the amplification of a single house keeping gene which expression could slightly change in one cell type compared to the other. As in the expression profiles generated from the DNA microarray dataset, RAC2, PAK1 and CFL2 from the Fc Gamma R-mediated phagocytosis showed an upregulation in the metastatic phenotype. Conversely, VAV2 and LIMK1 were supposed to respectively display a down- and an upregulation in the MDA-MB-231 cell type according to the expression profiles generated from the DNA microarray dataset, but real-time RT-PCR results did not confirm these expectations (Figure 5). For the genes of the toll like receptor signalling pathway, only IRF3 displayed a downregulation in the metastatic phenotype as predicted by the *in silico* expression profile. Indeed, the transcript level of TICAM1, TRAF3 and TBK1 was higher in MDA-MB-231 cells than in MCF-7 cells (Figure 6), while expression profiles generated from the DNA microarray dataset indicate the opposite results. This discrepancy between the results of the two techniques probably derives from the fact that the DNA microarray experiment and the real-time RT-PCR were achieved in different labs with conditions and cells different enough to cause such a discrepancy for some genes. Moreover, it should be noted that even for the genes displaying the same regulation in the DNA microarray and in the real-time RT-PCR expression profiles, the fold

inductions between the two cell types are very different between the two techniques. Again, this probably derives from the fact that the DNA microarray experiment and the real-time RT-PCR were achieved in different labs with different conditions and different cells.

However, the transcription of several genes (RAC2, PAK1, CFL2 and IRF3) displays the same regulation. Three of them (RAC2, PAK1 and CFL2) belong to the same pathway. That is why we chose to investigate their role in the migration of cancer cells. Since each of these genes are upregulated in the metastatic phenotype, they have been invalidated in the highly metastatic MDA-MB-231 cells to see if the cells are still able to migrate without the expression of these genes. The results showed that when PAK1 or CFL2 was invalidated, MDA-MB-231 cells showed an important decrease in migration (Figures 9 and 10). We did not succeed to achieve invalidation for RAC2 (data not shown). However, its role in the cancer cell migration should be further investigated.

RAC2 encodes a small GTPase of the RAS superfamily of proteins. It is known to be involved in the regulation of many cellular processes linked to the cell cycle [23]. Indeed, RAC2 controls the cell growth, the cytoskeletal reorganization and the activation of protein kinases. Therefore, it is not surprising that RAC2 has already been shown to be involved in the development of cancer, especially leukemia [24]. The protein encoded by PAK1 is a target of RAC2. PAK1 codes for a kinase. Like RAC2, PAK1 is involved in cytoskeleton reorganization. It is also involved in the nuclear signalling, cell motility and the regulation of cell morphology. Very recent studies have indicated a possible role for PAK1 in the metastasis of cancer cells [25, 26]. CFL2 encodes a component of the cytoskeleton. It allows the polymerization and the depolymerization of actin [27]. Although the mutation of CFL2 causes a myopathy [28], this gene has never been described as involved in metastasis or cancer.

To ensure that the decrease of migration observed during the scratch assays was not due to a decrease in proliferation due to the siRNA transfection, an MTT test was performed. A decrease in cell migration was observed 48 h after the siRNA transfection. However, we observed a similar decrease in cell migration when the siRNA targeting PAK1 or CFL2 or when the non-targeting siRNA were used. Since cell migration did not vary when the non-targeting siRNA was used, this suggets that the decrease in migration is not due to the decrease in cell proliferation.

Taken together, these results indicate a clear involvement of PAK1 and CFL2 in the metastatic phenotype of cancer cells. *In vivo* validations should be the next step in the

investigation of the role of these genes in the cancer cell migration process. MDA-MB-231 cells with PAK1 or CFL2 invalidated or MCF-7 cells with PAK1 or CFL2 overexpressed should be injected to mice to check if migration capacity is lost or gained, respectively. These results may provide new targets for therapy.

Conclusion

Based on the results of a previously published DNA microarray meta-analysis methodology, we performed several *in vitro* experiments aimed at identifying new genes involved in cancer metastasis. These assays aimed to demonstrate the possible implication of pathogen recognition and phagocytosis pathways in the metastatic phenotype of cancer cells. *In silico* expression profiles have been generated for 9 genes from 2 pathways. Real-time PCR confirmed these expression profiles for 4 genes. Two genes of the Fc Gamma R-mediated phagocytosis (PAK1 and CFL2) were invalidated in highly metastatic cancer cells to confirm their involvement in the cancer cell migration. Cell migration was strongly decreased when one of these two genes was invalidated. Moreover, a MTT test showed that the decrease in migration was not due to a decrease in proliferation. We therefore propose that PAK1 and CFL2 could be new targets for therapy for metastatic cancers.

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

MP carried out all the experiments, AN helped in the culture of MDA-MB-231 and MCF-7 cells, LL helped to carry the MTT assay, KV helped to carry the western blot, ED and CM conceived the study, participated in its design and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

M. Pierre is supported by FRIA (Belgium).

References

1. Friedl P, Wolf K: **Tumour-cell invasion and migration: diversity and escape mechanisms**. *Nat Rev Cancer* 2003, **3**(5):362-374.
2. Pantel K, Brakenhoff RH: **Dissecting the metastatic cascade**. *Nat Rev Cancer* 2004, **4**(6):448-456.
3. Chan DA, Giaccia AJ: **Hypoxia, gene expression, and metastasis**. *Cancer Metastasis Rev* 2007, **26**(2):333-339.
4. Gordan JD, Simon MC: **Hypoxia-inducible factors: central regulators of the tumor phenotype**. *Curr Opin Genet Dev* 2007, **17**(1):71-77.
5. Vaupel P: **The role of hypoxia-induced factors in tumor progression**. *Oncologist* 2004, **9 Suppl 5**:10-17.
6. Sullivan R, Graham CH: **Hypoxia-driven selection of the metastatic phenotype**. *Cancer Metastasis Rev* 2007, **26**(2):319-331.
7. Pierre M, DeHertogh B, DeMeulder B, Bareke E, Depiereux S, Michiels C, Depiereux E: **Enhanced Meta-analysis Highlights Genes Involved in Metastasis from Several Microarray Datasets**. *Journal of Proteomics and Bioinformatics* 2011, **4**(2):036-043.
8. Pierre M, DeHertogh B, Gaigneaux A, DeMeulder B, Berger F, Bareke E, Michiels C, Depiereux E: **Meta-analysis of archived DNA microarrays identifies genes regulated by hypoxia and involved in a metastatic phenotype in cancer cells**. *BMC Cancer* 2010, **10**:176.
9. Kronick MN: **Creation of the whole human genome microarray**. *Expert Rev Proteomics* 2004, **1**(1):19-28.
10. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society* 1995, **57**:289-300.
11. Gautier L, Moller M, Friis-Hansen L, Knudsen S: **Alternative mapping of probes to genes for Affymetrix chips**. *BMC Bioinformatics* 2004, **5**:111.
12. Gur-Dedeoglu B, Konu O, Kir S, Ozturk AR, Bozkurt B, Ergul G, Yulug IG: **A resampling-based meta-analysis for detection of differential gene expression in breast cancer**. *BMC Cancer* 2008, **8**:396.
13. Ma S, Huang J: **Regularized gene selection in cancer microarray meta-analysis**. *BMC Bioinformatics* 2009, **10**:1.
14. Ochsner SA, Steffen DL, Hilsenbeck SG, Chen ES, Watkins C, McKenna NJ: **GEMS (Gene Expression MetaSignatures), a Web resource for querying meta-analysis of expression microarray datasets: 17beta-estradiol in MCF-7 cells**. *Cancer Res* 2009, **69**(1):23-26.
15. NCBI: **Gene Expression Omnibus**. In.; 2000.
16. Liu H, Zeeberg BR, Qu G, Koru AG, Ferrucci A, Kahn A, Ryan MC, Nuhanovic A, Munson PJ, Reinhold WC *et al*: **AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets**. *Bioinformatics* 2007, **23**(18):2385-2390.

17. Wu Z, Irizarry R, Gentleman R, Murillo F, Spencer F: **A model-based background adjustment for oligonucleotide expression arrays**. *Journal of the American Statistical Association* 2004, **99**:909-917.

18. Pfaffl MW: **A new mathematical model for relative quantification in real-time RT-PCR**. *Nucleic Acids Res* 2001, **29**(9):e45.

19. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Res* 1999, **27**(1):29-34.

20. Groves E, Dart AE, Covarelli V, Caron E: **Molecular mechanisms of phagocytic uptake in mammalian cells**. *Cell Mol Life Sci* 2008, **65**(13):1957-1976.

21. May RC, Machesky LM: **Phagocytosis and the actin cytoskeleton**. *J Cell Sci* 2001, **114**(Pt 6):1061-1077.

22. Kawai T, Akira S: **Antiviral signaling through pattern recognition receptors**. *J Biochem* 2007, **141**(2):137-145.

23. Pai SY, Kim C, Williams DA: **Rac GTPases in human diseases**. *Dis Markers* 2010, **29**(3-4):177-187.

24. Holland M, Castro FV, Alexander S, Smith D, Liu J, Walker M, Bitton D, Mulryan K, Ashton G, Blaylock M *et al*: **RAC2, AEP, and ICAM1 expression are associated with CNS disease in a mouse model of pre-B childhood acute lymphoblastic leukemia**. *Blood* 2011.

25. Huynh N, Liu KH, Baldwin GS, He H: **P21-activated kinase 1 stimulates colon cancer cell growth and migration/invasion via ERK- and AKT-dependent pathways**. *Biochim Biophys Acta* 2010, **1803**(9):1106-1113.

26. Li LH, Zheng MH, Luo Q, Ye Q, Feng B, Lu AG, Wang ML, Chen XH, Su LP, Liu BY: **P21-activated protein kinase 1 induces colorectal cancer metastasis involving ERK activation and phosphorylation of FAK at Ser-910**. *Int J Oncol* 2010, **37**(4):951-962.

27. Maciver SK, Hussey PJ: **The ADF/cofilin family: actin-remodeling proteins**. *Genome Biol* 2002, **3**(5):reviews3007.

28. Agrawal PB, Greenleaf RS, Tomczak KK, Lehtokari VL, Wallgren-Pettersson C, Wallefeld W, Laing NG, Darras BT, Maciver SK, Dormitzer PR *et al*: **Nemaline myopathy with minicores caused by mutation of the CFL2 gene encoding the skeletal muscle actin-binding protein, cofilin-2**. *Am J Hum Genet* 2007, **80**(1):162-167.

Figures

Figure 1 – Fc Gamma R-mediated phagocytosis pathway

The Fc Gamma R-mediated phagocytosis pathway was retrieved from the KEGG database. The expression profiles of the genes of the sub-pathway in red are presented in Figure 3.

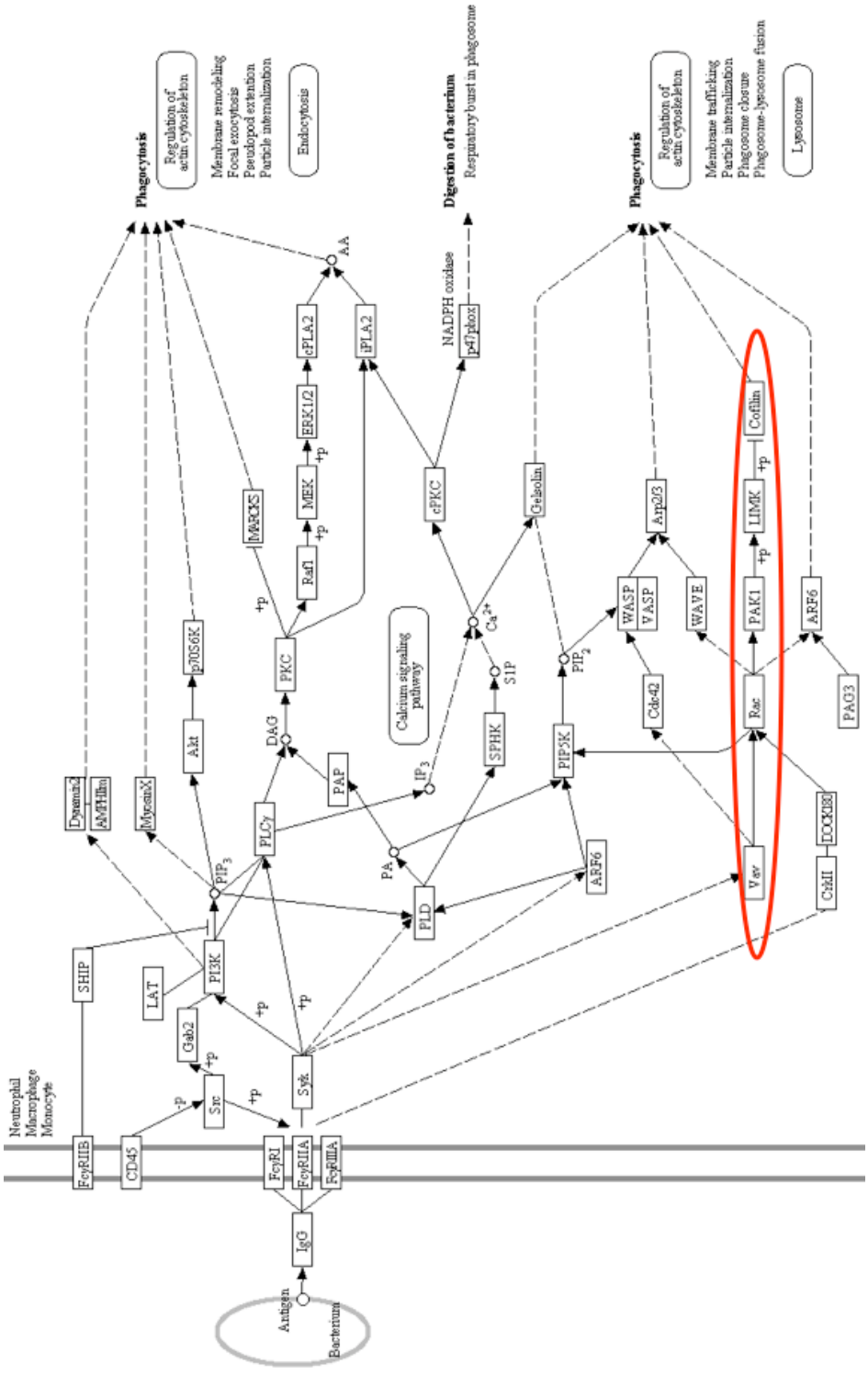


Figure 2 – Toll like receptor signalling pathway

The toll like receptor signalling pathway was retrieved from the KEGG database. The expression profiles of the genes of the sub-pathway in red are presented in Figure 4.

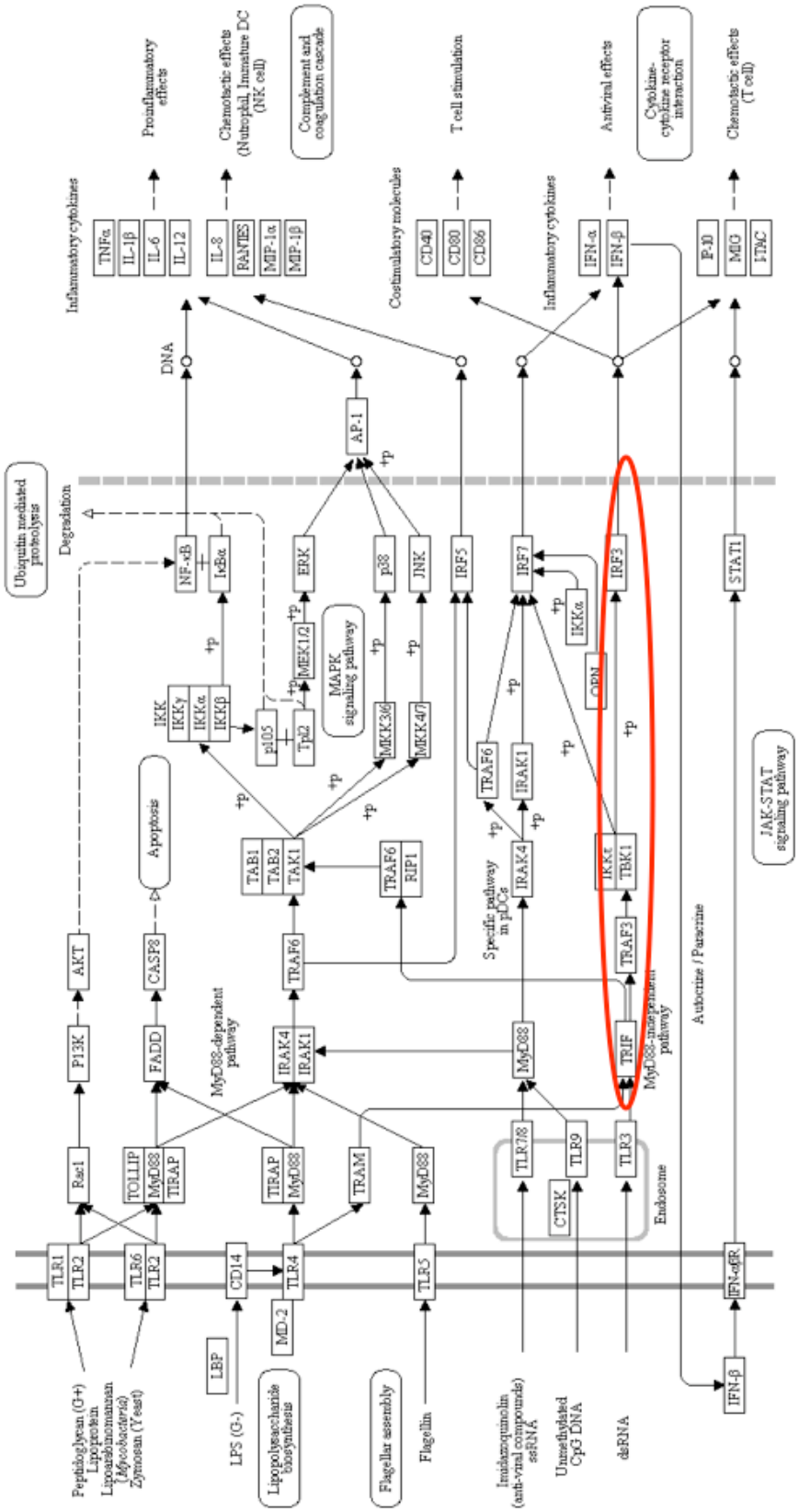


Figure 3 – Expression profiles of genes involved in the Fc Gamma R-mediated phagocytosis pathway

The X axis shows the two cell types compared: MCF-7 and MDA-MB-231. The Y axis shows the expression value that reflects the transcript level of the (A) VAV2, (B) RAC2, (C) PAK1, (D) LIMK1 and (E) CFL2 genes in each sample. The links between the dots do not represent any relation between the samples.

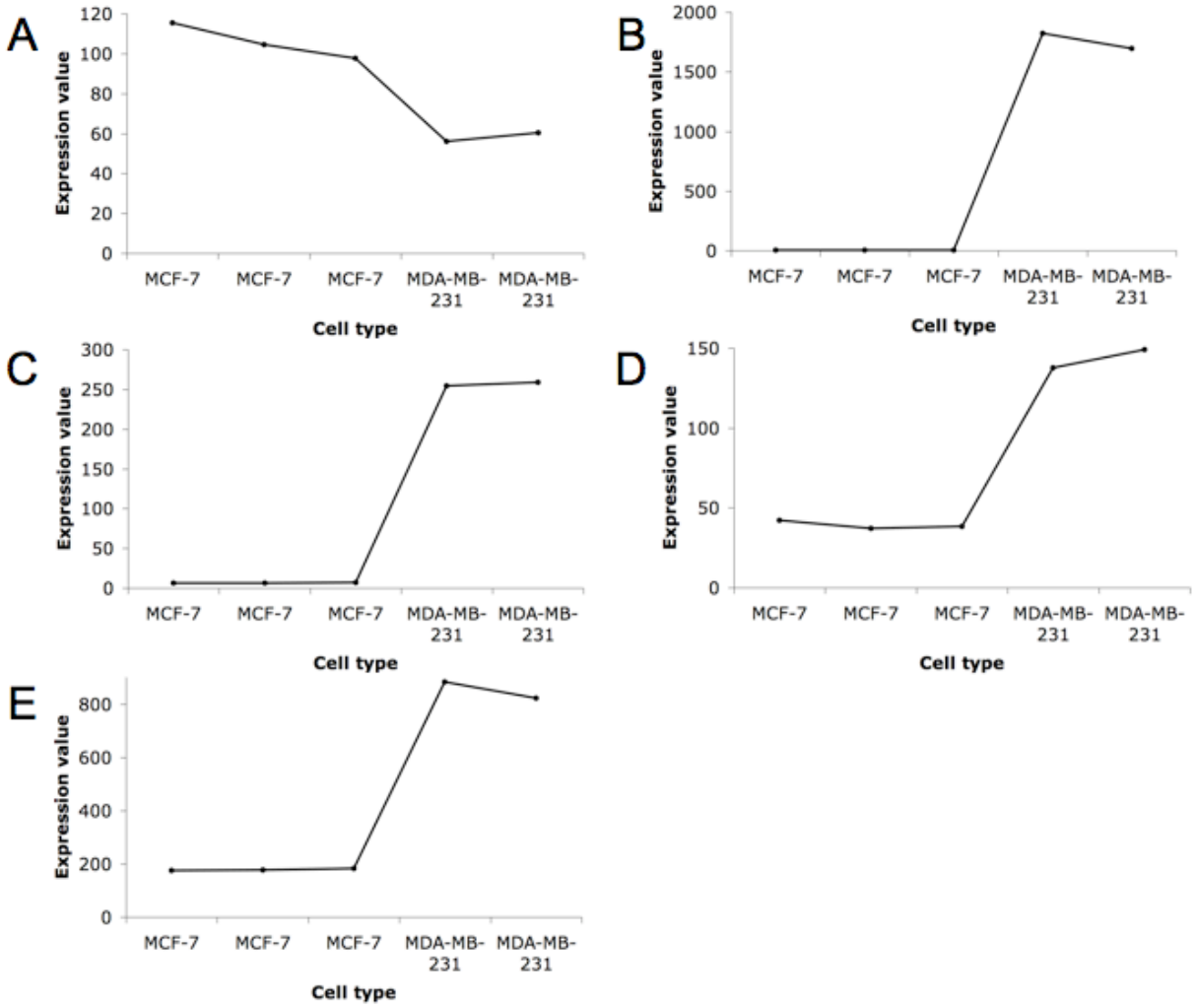


Figure 4 – Expression profiles of genes involved in the toll like receptor signalling pathway

The X axis shows the two cell types compared: MCF-7 and MDA-MB-231. The Y axis shows the expression value that reflects the transcript level of the (A) TICAM1, (B) TRAF3, (C) TBK1 and (D) IRF3 genes in each sample. The links between the dots do not represent any relation between the samples.

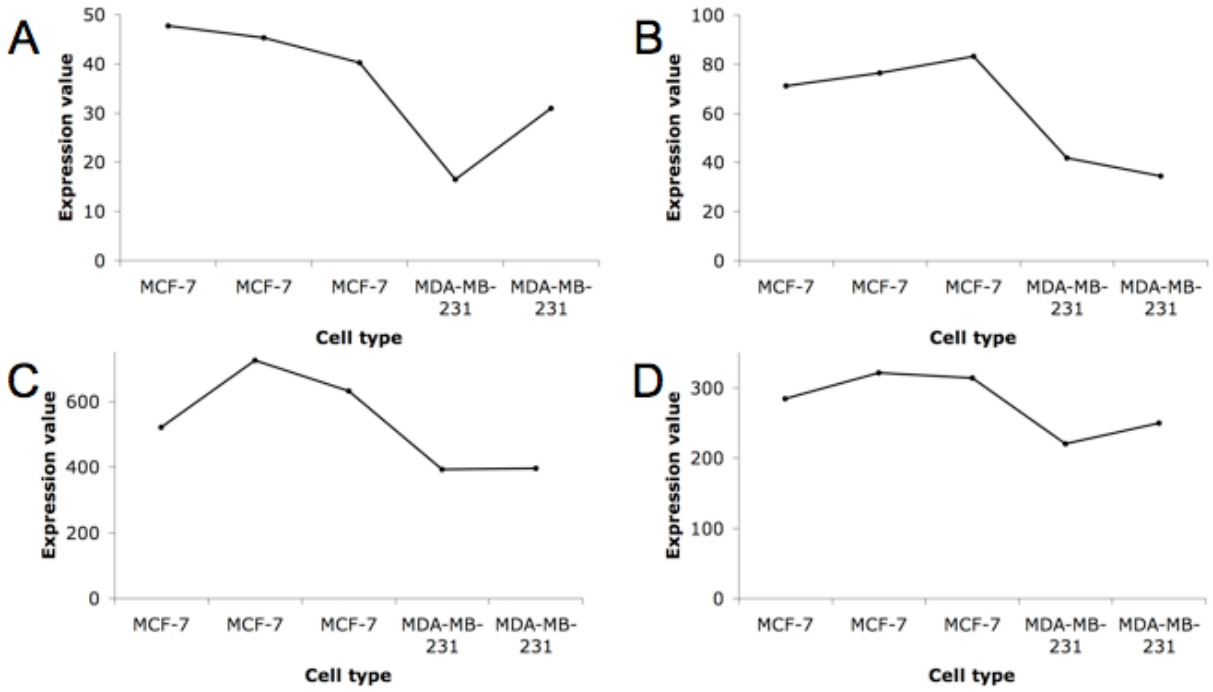


Figure 5 – Results of the real-time RT-PCR of the genes involved in the Fc Gamma R-mediated phagocytosis pathway

The mean amplification of 23kDa and α-tubulin was computed and used as endogenous standards to allow the comparison of two different cell types. The results for VAV2, RAC2, PAK1, LIMK1 and CFL2 are expressed in fold induction for three independent experiments, as means ± SD (n = 3), in MDA-MB-231 cells compared to MCF-7 cells. An ANOVA was performed for each gene to assess the statistical significance between the MDA-MB-231 and the MCF-7 cell types. NS = non significant difference ($P<0.05$) between the delta Cts in MDA-MB-231 cells and the delta Cts in MCF-7 cells. S = significant difference ($P<0.05$) between the delta Cts in MDA-MB-231 cells and the delta Cts in MCF-7 cells. VHS = very highly significant difference ($P<0.001$) between the delta Cts in MDA-MB-231 cells and the delta Cts in MCF-7 cells. VH = the ANOVA was impossible to perform because of variance heterogeneity.

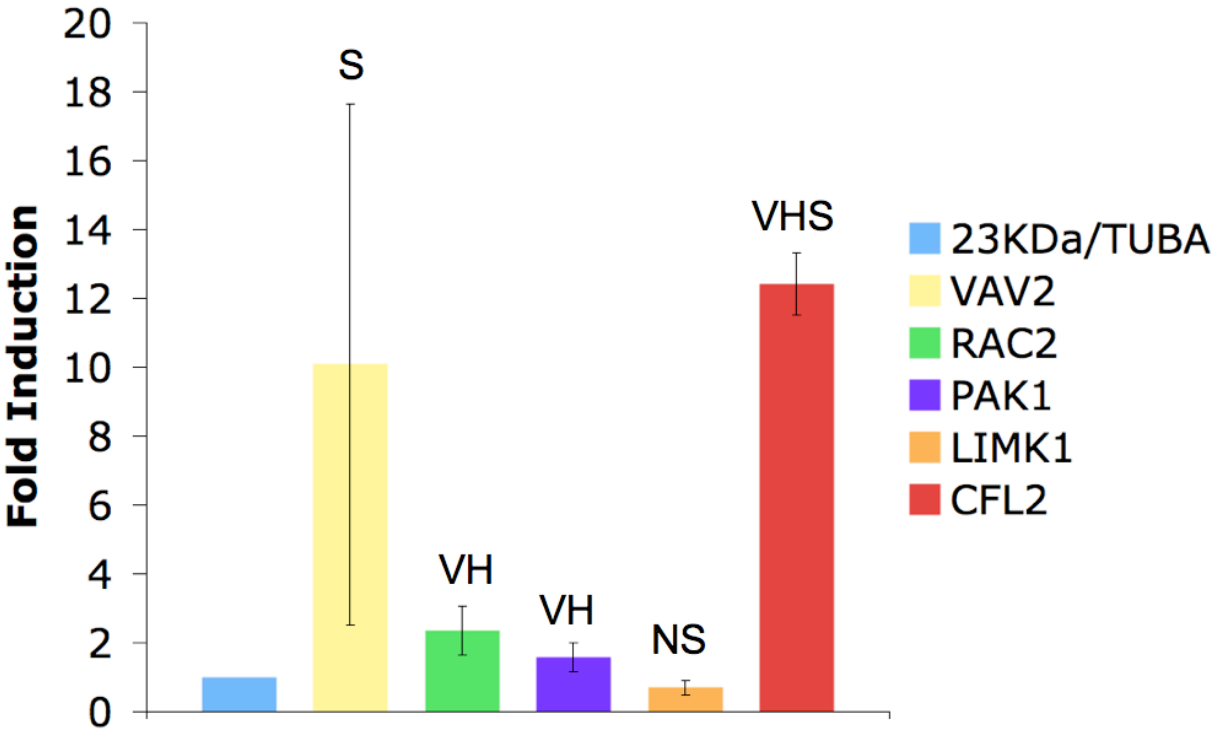


Figure 6 – Results of the real-time RT-PCR of the genes involved in the toll like receptor signalling pathway

The mean amplification of 23kDa and α-tubulin was computed and used as endogenous standards to allow the comparison of two different cell types. The results for TICAM1, TRAF3, TBK1 and IRF3 are expressed in fold induction for three independent experiments, as means ± SD (n = 3), in MDA-MB-231 cells compared to MCF-7 cells. An ANOVA was performed for each gene to assess the statistical significance between the MDA-MB-231 and the MCF-7 cell types. NS = non significant difference (P<0.05) between the delta Cts in MDA-MB-231 cells and the delta Cts in MCF-7 cells. HS = highly significant difference (P<0.01) between the delta Cts in MDA-MB-231 cells and the delta Cts in MCF-7 cells. VHS = very highly significant difference (P<0.001) between the delta Cts in MDA-MB-231 cells and the delta Cts in MCF-7 cells. VH = the ANOVA was impossible to perform because of variance heterogeneity.

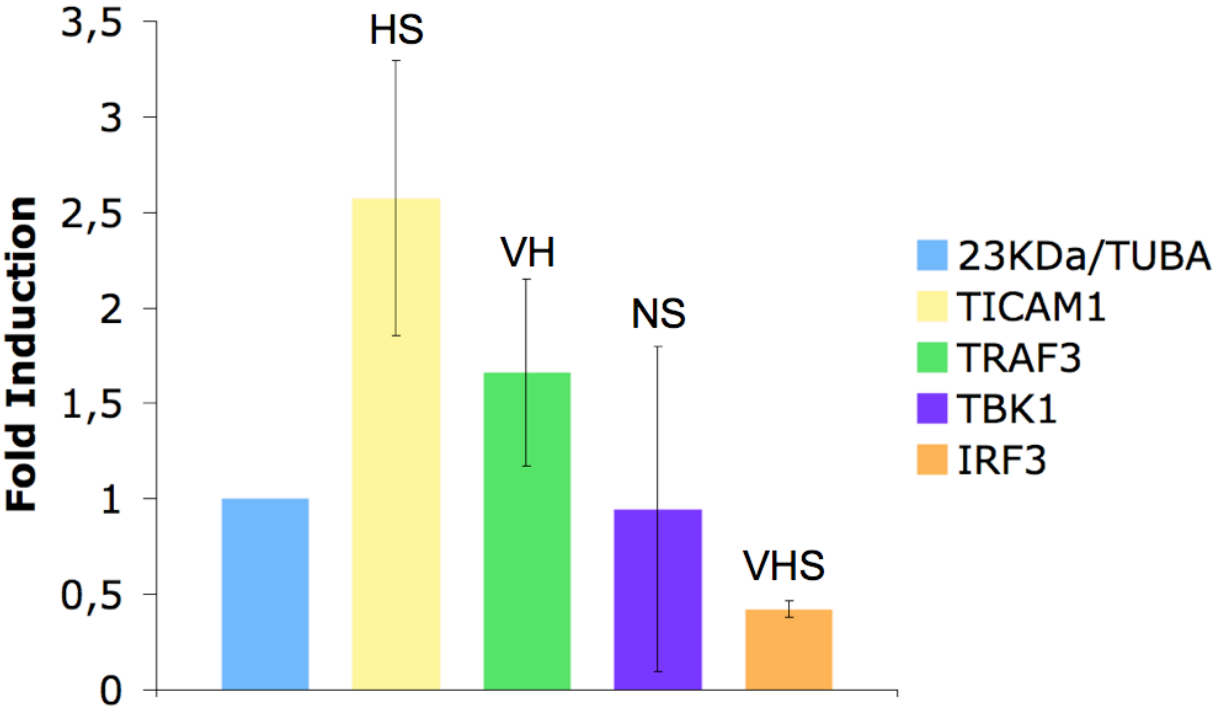


Figure 7 – Transcript levels of genes in siRNA transfected cells

α-tubulin was used as the endogenous standard. The results for PAK1 and CFL2 are expressed in fold induction for three independent experiments, as means ± SD (n = 3), in MDA-MB-231 cells transfected with non-targeting siRNA (A), PAK1 siRNA (B) or CFL2 siRNA (C) compared to control MDA-MB-231 cells.

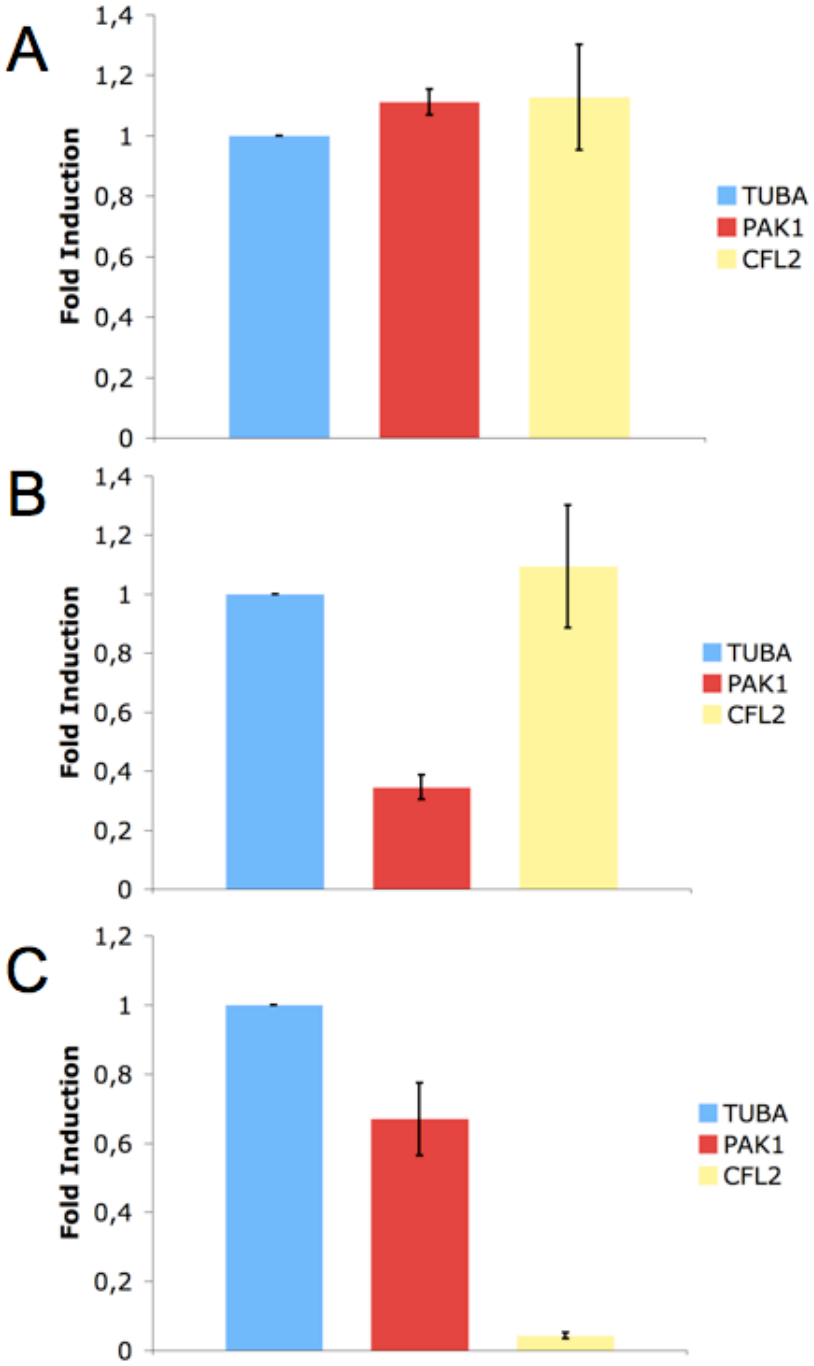


Figure 8 – Protein levels of siRNA transfected cells

MDA-MB-231 cells were transfected with non-targeting siRNA (NT), PAK1 siRNA or CFL2 siRNA. Total cell lysates were run on SDS-PAGE (10%) and probed with an anti-PAK1 antibody. Control MDA-MB-231 cells were also assayed.

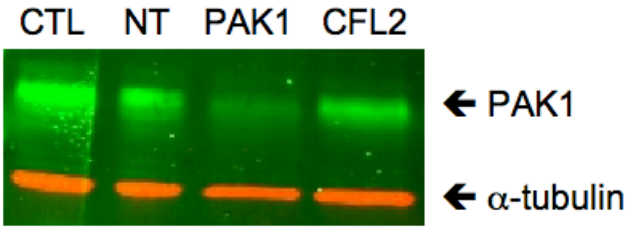


Figure 9 – Scratch assays

MDA-MB-231 cells were transfected with non-targeting siRNA (NT), PAK1 siRNA or CFL2 siRNA. Small scratches were made in the cell monolayer. Pictures of the scratches were taken at 0 h, 6 h and 12 h after they have been made. Control MDA-MB-231 cells were also assayed.

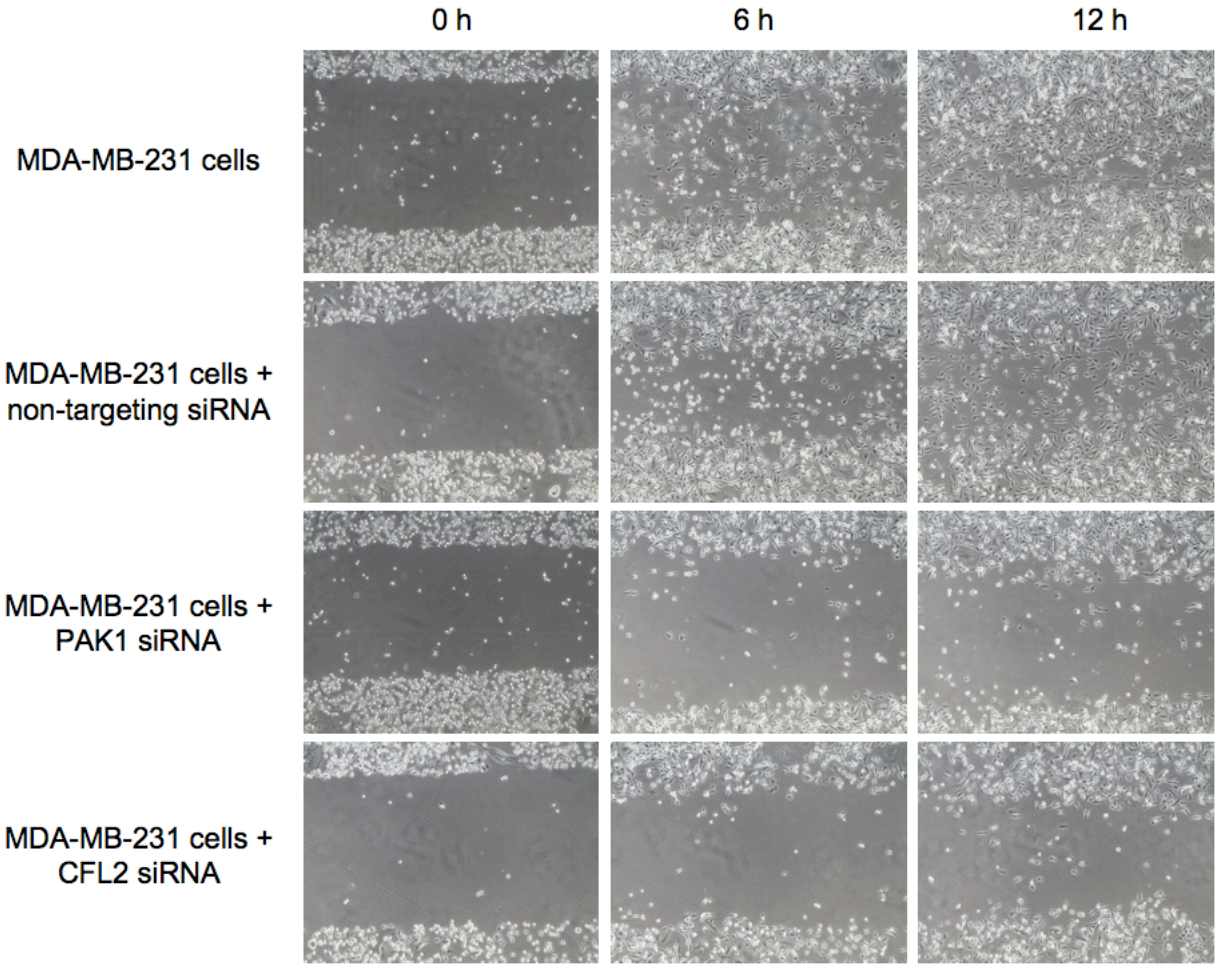


Figure 10 – Quantification of cell migration

The results are expressed in number of cells present in the scratch for three independent experiments, as means ± SD (n = 3), 6 h (in red) and 12 h (in yellow) after it has been made for control MDA-MB-231 cells (A), for MDA-MB-231 cells transfected with non-targeting siRNA (B), for MDA-MB-231 cells transfected with PAK1 siRNA (C) and for MDA-MB-231 cells transfected with CFL2 siRNA (D).

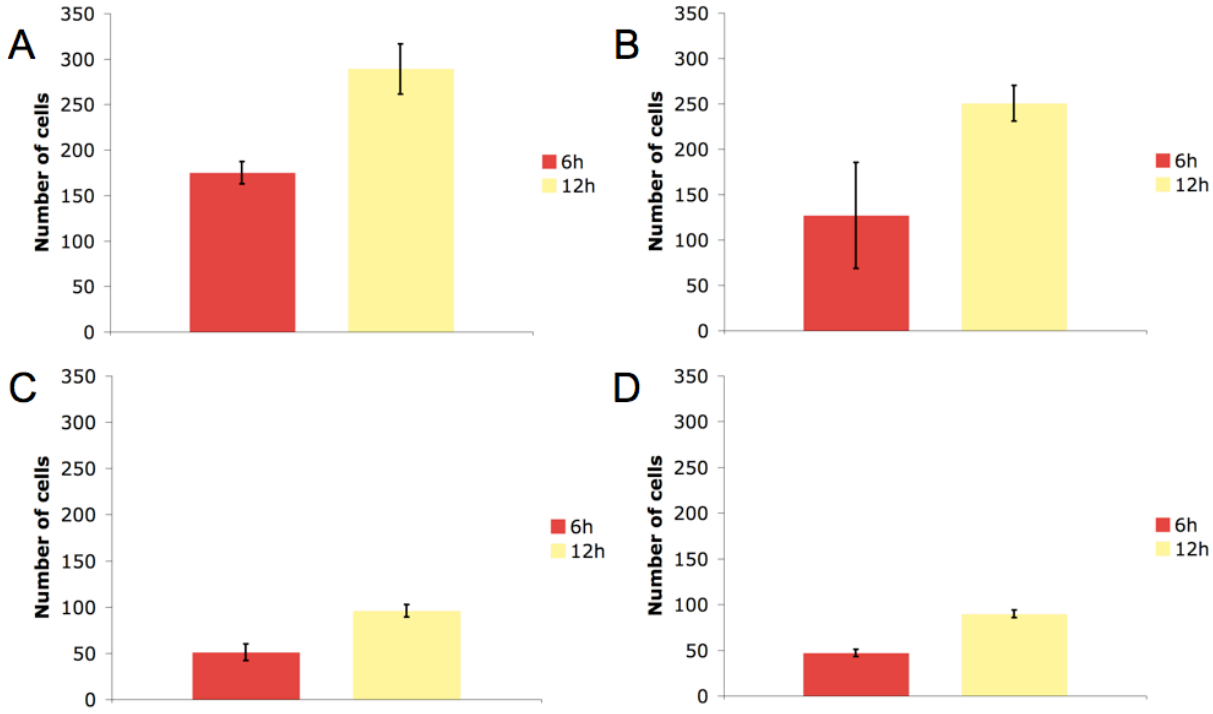
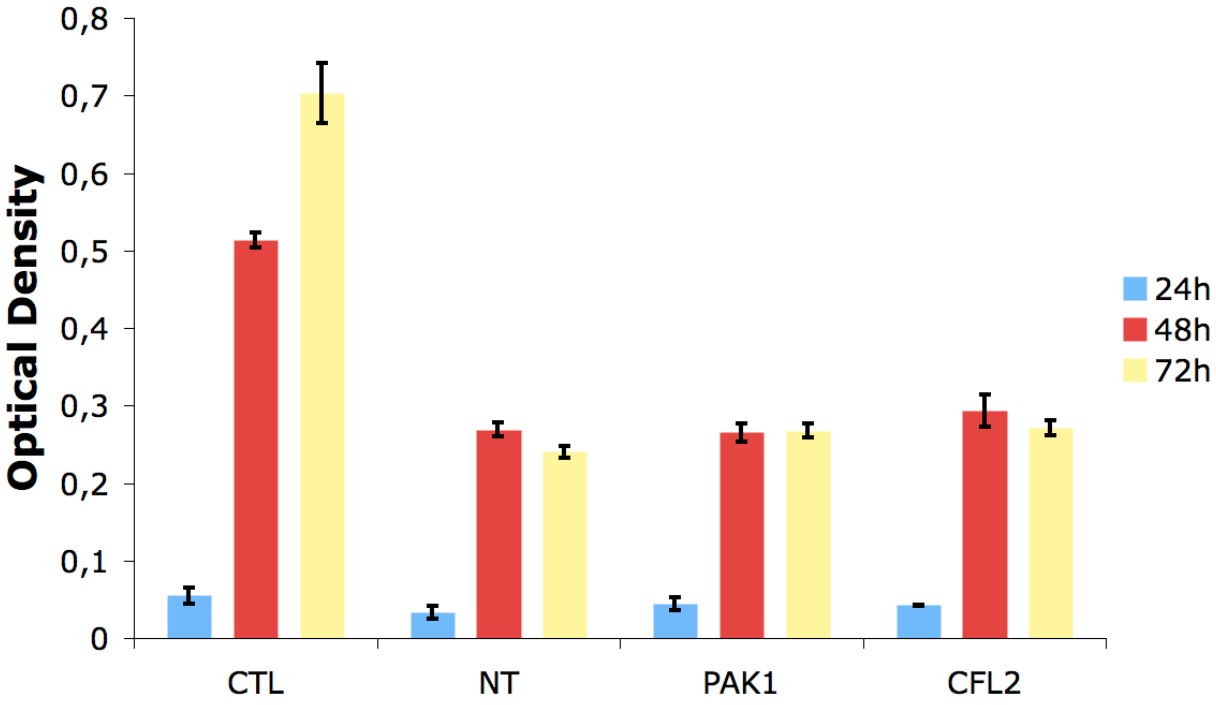


Figure 11 – Cell viability after siRNA transfection

The X axis shows the siRNA used on the MDA-MB-231 cell cultures (CTL = no siRNA, NT = non-targeting siRNA, RAC2 = RAC2 siRNA, PAK1 = PAK1 siRNA and CFL2 = CFL2 siRNA) and the time after the cells have been transfected (24 h, 48 h and 72 h). The Y axis shows the optical density of the cells at 570 nm. The results are expressed in optical density for three independent experiments, as means ± SD (n = 3).



Additional files

Additional file 1 – Expression profiles of genes involved in the Fc Gamma R-mediated phagocytosis pathway

The X axis shows the two cell types compared: MCF-7 and MDA-MB-231. The Y axis shows the expression value that reflects the transcript level of the genes in the sample.

Additional file 2 – Expression profiles of genes involved in the toll like receptor signalling pathway

The X axis shows the two cell types compared: MCF-7 and MDA-MB-231. The Y axis shows the expression value that reflects the transcript level of the genes in the sample.

IV. DISCUSSION & CONCLUSION

De nombreuses études tentent aujourd'hui de comprendre les mécanismes qui mènent au phénotype métastatique. Dans cette optique, ce projet avait pour objectif d'exploiter l'importante quantité de données venant des expériences utilisant des puces à ADN et traitant des métastases pour générer des hypothèses quant à l'implication de gènes dans le processus métastatique. Toute hypothèse, n'ayant pas déjà fait l'objet de recherches approfondies de la part de la communauté scientifique, pourrait alors être testée par des techniques expérimentales afin de la valider. De cette façon, nous pourrions fournir de nouvelles cibles pour de futures thérapies contre cette maladie. Cependant, tant au niveau de la méthodologie bioinformatique que nous proposons, que des validations expérimentales que nous avons réalisées, des choix différents auraient pu être faits et des questions restent en suspens. Cette partie va discuter de ces problèmes et des perspectives laissées par ce travail.

La première partie du projet impliquait de sélectionner une série de jeux de données de puces à ADN. Ceux-ci venaient des deux bases de données publiques les plus importantes : Gene Expression Omnibus et ArrayExpress [150]. Nous avons choisi ces deux bases de données car elles sont devenues les lieux de dépôt obligatoires de tout jeu de données de puces à ADN pour quiconque souhaite publier un article dans lequel des puces à ADN ont été utilisées. En effet, aujourd'hui, toutes les revues scientifiques demandent aux auteurs de déposer leurs jeux de données sur l'une de ces bases de données afin de permettre leur ré-analyse éventuelle. L'avantage d'un tel système est que les jeux de données candidats à une publication sont aussi disponibles pour les critiques des revues scientifiques. Ainsi, en plus d'être soumis aux normes exigées par les bases de données, ces jeux de données sont, en plus, examinés et validés par les critiques des revues scientifiques. Cependant, tous les jeux de données déposés dans les bases de données ne font pas l'objet d'une publication. Ces jeux de données sont susceptibles de générer des résultats erronés. Nous avons vérifié que tous les jeux de données utilisés dans la présente étude avaient fait l'objet d'une publication.

Il faut aussi signaler qu'en 2006 est né le projet MicroArray Quality Control (MAQC) [209]. Celui-ci a pour but de développer des standards afin de contrôler la qualité des jeux de données générés à partir de puces à ADN tant au niveau de la conception de celles-ci, qu'au niveau de l'extraction de l'ARN ou de l'ADN et de son hybridation aux puces, et qu'au niveau de l'analyse des données générées. Pour cela, les protocoles expérimentaux ont été comparés ainsi que les différents modèles et marques de puces à ADN et que les différentes méthodes d'analyse. Il en ressort toute une série de lignes de conduite à respecter lors de la réalisation et de l'analyse d'expériences utilisant des puces à ADN. De plus, les conclusions

du projet, qui se poursuit toujours aujourd'hui [210, 211], sont que les résultats générés par l'analyse de puces à ADN nécessitent d'être validés par d'autres approches expérimentales comme nous l'avons fait au cours de ce travail.

Ensuite, nous avons choisi de ne sélectionner que des puces à ADN de la marque Affymetrix. Ce choix a été réalisé d'abord parce que plusieurs projets du laboratoire ont été menés sur cette technologie, nous disposons donc d'outils qui sont spécifiques à l'analyse de ce type de puce ; et comme le but de ce projet n'était pas d'étendre l'utilisation de ces outils à d'autres technologies, nous nous sommes volontairement limités aux GeneChips Affymetrix. Nous avons également choisi cette technologie car elle est la plus répandue et nous donnait accès à une masse d'informations suffisante pour réaliser le projet. Il est d'ailleurs intéressant de noter qu'à la vitesse à laquelle s'accumulent les jeux de données sur les bases de données, une nouvelle recherche menée aujourd'hui apporterait de nouveaux jeux de données qui n'ont pas été utilisés dans la présente étude. En effet, les jeux de données qui ont été utilisés ici ont été rapatriés des bases de données en 2007. En 2010, nous avons rapatrié plusieurs jeux de données qui n'étaient pas présents trois ans auparavant afin de valider les premières hypothèses que nous avions générées. Ceci montre l'utilité d'outils comme PathEx, qui a été développé au sein de notre laboratoire en 2010, dont l'une des fonctions est de rapatrier des jeux de données, en fonction de critères spécifiques, à partir de Gene Expression Omnibus et d'ArrayExpress [197]. De plus, comme la mise à jour de PathEx est régulière, il est aisé de rapatrier les jeux de données les plus récents.

Comme ce travail avait pour objectif de mettre en évidence des gènes impliqués dans la régulation du potentiel métastatique, nous avons sélectionné toute une série de jeux de données générés à partir de puces à ADN comparant des métastases aux tumeurs primaires à partir desquelles elles se sont détachées. Cependant, la question de savoir si des cellules ayant déjà métastasé avec succès présentent encore les niveaux de transcrits qui leur ont permis de se détacher de la tumeur primaire et de migrer pourrait être posée. En effet, cette étude tente de mettre en avant une expression différentielle de certains gènes entre les cellules métastatiques et non métastatiques. La validation de ces gènes par des techniques expérimentales indiquerait une possible implication de ces gènes dans le phénotype métastatique. Cependant, il est possible que les cellules qui entament la cascade métastatique sous- ou surexpriment des gènes de manière séquentielle en fonction de l'étape de la cascade à laquelle elles se trouvent. Une fois, l'étape franchie, l'expression des gènes ayant permis ce passage pourrait revenir à un taux similaire à celui des cellules non métastatiques. Dès lors,

des cellules qui ont terminé de métastaser et qui se sont implantées dans un site secondaire ne présenteraient plus d'expression différentielle de leurs gènes par rapport aux cellules non métastatiques. Cependant, de nombreuses études similaires à celle que nous présentons mais à plus petite échelle sont parvenues à mettre en évidence des gènes participant au développement métastatique en comparant des tumeurs primaires à des métastases bien établies. De plus, en admettant que les cellules des métastases n'expriment plus les gènes qui leur ont permis de métastaser, l'expression de ces gènes sera retrouvée dans les cellules des tumeurs primaires puisque ce sont celles-ci qui ont engendré les métastases et qu'elles ont dû, pour cela, sous- ou surexprimer des gènes spécifiques. Dans ce cas, l'ARN provenant des cellules de la tumeur primaire hybridé sur les puces à ADN serait un mélange d'ARN de cellules sous- ou surexprimant les gènes nécessaires à la progression dans la cascade métastatique et de cellules qui ne mettent pas en place cette cascade. Finalement, lors de l'analyse des données venant des puces à ADN, cela résulterait en un signal moins intense, mais pas nul. Nous pensons donc que notre approche peut permettre de mettre en évidence des gènes dont l'expression est régulée en fonction de leur potentiel métastatique.

De la même manière, l'on pourrait se demander à quel point les cellules non cancéreuses se trouvant au sein des tumeurs influent sur les résultats générés à partir d'expériences utilisant des puces à ADN. En effet, pour la réalisation d'une telle expérience, une biopsie est prélevée, celle-ci contient à la fois des cellules cancéreuses mais également des cellules non cancéreuses, dites stromales (cellules saines du tissu comme les fibroblastes ou les cellules du système immunitaire). Ensuite, l'ARN de cette biopsie est extrait et hybridé sur les puces à ADN (après plusieurs étapes expérimentales). Comme au cours de ce travail nous avons tenté de mettre en évidence une expression différentielle des gènes entre des tumeurs métastatiques et non métastatiques, le fait que l'ARN de cellules non cancéreuses soit aussi présent sur les puces à ADN où se trouvent l'ARN des cellules non métastatiques pourrait poser problème. C'est pourquoi les gènes identifiés comme étant exprimés de manière différentielle entre les tumeurs métastatiques et non métastatiques par l'analyse des résultats générés par des puces à ADN doivent impérativement être validés expérimentalement.

Pour réaliser l'étude que nous présentons, nous avons analysé individuellement chaque jeu de données en utilisant les CDFs d'AffyProbeMiner plutôt que ceux d'Affymetrix. Nous avons opéré ce choix pour plusieurs raisons. Premièrement, comme AffyProbeMiner a été mis à disposition du public en 2007, il s'agissait de l'outil le plus récent proposant des CDFs

alternatifs. Ceci signifie que ces CDFs ont été conçus sur base de l'information la plus récente provenant des bases de données génomiques. Deuxièmement, AffyProbeMiner était le seul outil à proposer des CDFs alternatifs pour tous les modèles de GeneChip Affymetrix. Notre étude utilisant huit modèles différents de GeneChips, il était nécessaire de disposer d'un CDF alternatif pour chacun de ces modèles. Enfin, AffyProbeMiner, en plus de proposer des CDFs alternatifs, met à disposition les scripts permettant de les exploiter, ce qui en fait un outil prêt à l'emploi.

Nous avons montré (Figures 23 et 24) que l'utilisation des CDFs d'AffyProbeMiner changeait substantiellement les résultats d'une analyse. Afin de démontrer que les CDFs d'AffyProbeMiner sont meilleurs, il faudrait les utiliser sur des jeux de données où la quantité d'ARN, correspondant à certains gènes, est connue et contrôlée, comme les carrés latins d'Affymetrix. Cependant, les gènes dont la quantité d'ARN est contrôlée sur les carrés latins d'Affymetrix sont peu nombreux (42 gènes pour 22.000 représentés sur la puce) et sont des gènes qui étaient déjà bien connus quand les GeneChips ont été commercialisés, leur séquence n'a donc pas évolué dans les bases de données génomiques. Les mêmes sondes sont donc attribuées à ces gènes, que ce soit avec les CDFs alternatifs ou avec les CDFs standards. Ce n'est donc pas pour la mesure de l'expression de ces gènes que l'utilisation de CDFs alternatifs peut se révéler utile. Il faudrait, comme l'ont fait Choe et ses collègues [191] pour la drosophile, générer un jeu de données où les gènes, dont la quantité d'ARN est contrôlée, sont suffisamment nombreux pour inclure des gènes qui ont vu leur séquence évoluer au sein des bases de données génomiques.

Cependant, même si à l'heure actuelle, un tel jeu de données n'existe pas pour l'être humain, un argument en faveur des CDFs alternatifs vient du fait que leur utilisation sur des jeux de données traitant de processus biologiques bien connus permet d'obtenir des résultats plus cohérents qu'avec les CDFs standards. C'est pourquoi nous recommandons l'utilisation de CDFs alternatifs lors de l'analyse de puces à ADN. Cependant, comme AffyProbeMiner a été rendu disponible en 2007, il est fort probable que les séquences et les annotations des gènes aient encore évolué depuis lors. Pour toute nouvelle utilisation, il serait donc utile de mettre à jour les CDFs afin que ceux-ci soient les plus cohérents possibles avec la connaissance actuelle du génome.

Pour le prétraitement des données, nous avons utilisé GCRMA car aucun benchmark n'a montré l'existence d'une meilleure méthode que celle-ci. De plus, les outils développés au

laboratoire étant optimisés pour fonctionner avec celui-ci, notre choix s'est naturellement porté sur GCRMA.

Pour le traitement statistique des données, nous avons choisi d'utiliser le Window t test [182], qui a été développé au sein du laboratoire. En 2010, un benchmark des méthodes de traitement statistique des données venant de puces à ADN [192] a montré qu'une méthode était légèrement supérieure au Window t test : le shrinkage t test [190]. L'utilisation de ce test pourrait donc encore améliorer les résultats. Toutefois, le Window t test conserve certains avantages par rapport au shrinkage t test. D'abord, il est implémenté dans le package Pegase, qui a été développé dans notre laboratoire, ce qui rend son utilisation plus simple. De plus, ce package permet d'associer au Window t test, une correction de Welch afin de parer aux éventuelles différences de variance entre les échantillons testés. Dans le cadre de ce projet, la correction de Welch a donc été appliquée. Enfin, les différences entre méthodes deviennent minimes lorsque le nombre de réplicats est supérieur à dix.

À la fin d'une analyse de données de puces à ADN, il est d'usage d'appliquer une correction pour tests multiples. La méthode la plus efficace à ce sujet est la correction proposée par Benjamini et Hochberg [194]. Cependant, son application rend les tests tellement stricts que le nombre de faux négatifs augmente dramatiquement, à tel point qu'il arrive que plus aucun gène ne soit sélectionné comme étant exprimé de manière différentielle entre deux conditions expérimentales. Afin de nous affranchir de ce seuil statistique rendu trop strict par une correction pour tests multiples, ou pas assez sans cette correction, les gènes ont été classés selon leur p value et nous nous sommes servi de ce classement pour les approches de méta-analyse.

Afin de s'assurer de la pertinence de chacune des approches utilisées dans la méta-analyse, nous avons compté le nombre de gènes connus comme étant impliqués dans les cancers, le phénotype métastatique ou la réponse à l'hypoxie dans les gènes rappatriés à la fois par les intersections et les intersections d'unions, par les intersections et les méta-analyses, ou par les intersections d'unions et les méta-analyses. Il en ressort que la combinaison des intersections et des intersections d'unions permet de rappatrier 65% de gènes qui ont déjà été décrits dans la littérature comme étant impliqués dans les cancers, le phénotype métastatique ou la réponse à l'hypoxie. Il en est de même pour 54% des gènes rappatriés par la combinaison des intersections et des méta-analyses et également pour 67% des gènes rappatriés par la combinaison des intersections d'unions et des méta-analyses. La contribution de chacune des approches à la méta-analyse est donc à peu près équivalente. Il

est aussi intéressant de noter que la combinaison des intersections et des méta-analyses ne permet de rappatrier que 3% de gènes connus comme étant impliqués dans la réponse à l'hypoxie contre 13% pour la combinaison des intersections et des intersections d'unions et 7% pour la combinaison des intersections d'unions et des méta-analyses. Bien sûr, cette observation n'est pas étonnante puisque les intersections d'unions ne sont pas prises en compte dans cette combinaison et qu'elles sont la seule approche où les jeux de données étudiant la réponse à l'hypoxie sont systématiquement intégrés.

La méthode de méta-analyse que nous proposons a donc permi de sélectionner 165 gènes d'intérêt. Avec l'outil DAVID, nous avons pu replacer ces 165 gènes dans 42 voies de signalisation différentes. Parmi celles-ci, 12 sont directement liées aux cancers, 5 sont liées à la prolifération et la mobilité cellulaire et 5 sont liées à la reconnaissance de pathogènes et à la phagocytose. Cependant, il faut noter que si autant de voies ont été rappatriées par DAVID, c'est notamment parce que certains gènes d'intérêt qui lui ont été soumis ont plusieurs fonctions biologiques et se retrouvent dès lors dans de nombreuses voies de signalisation. Par exemple, MAP2K1 est un gène d'intérêt sélectionné par la méta-analyse et qui apparaît dans 33 des 42 voies de signalisation rappatriées par DAVID, parmi lesquelles 11 des 12 voies liées aux cancers, les 5 voies liées à la prolifération et la mobilité cellulaire et les 5 liées à la reconnaissance de pathogènes et à la phagocytose. Afin de s'assurer que les voies de signalisation rappatriées par DAVID sont bien impliquées dans le développement métastatique et qu'il ne s'agit pas d'un biais dû à la présence de quelques gènes impliqués dans de nombreuses voies, il faut valider expérimentalement ces voies de signalisation comme nous avons tenté de le faire au cours de ce projet.

Avec l'expansion du nombre de jeux de données utilisant des puces à ADN rendus publiques, d'autres méthodes de méta-analyse ont vu le jour ces dernières années. Selon Cahan et ses collègues, celles-ci peuvent être de deux types : complètes ou comparatives [212]. Dans une méta-analyse complète, différents jeux de données sont rassemblés et une analyse complète est menée en une fois sur cet ensemble de jeux de données. Par contre, dans une méta-analyse comparative, les résultats générés individuellement par différents jeux de données sont comparés pour en tirer de nouvelles informations. Il n'y a donc pas de ré-analyse préalable comme c'est le cas pour les méta-analyses complètes, il s'agit d'une comparaison de listes de gènes.

Lors d'une méta-analyse complète, l'augmentation du nombre de réplicats permet d'augmenter la puissance statistique des tests et ainsi de diminuer le nombre de faux positifs

et de faux négatifs. Cependant, il faut rester conscient que, outre l'augmentation de variabilité inhérente à une telle pratique, seul environ un tiers des données brutes disponibles publiquement sont de qualité suffisante [213]. Ces observations n'ont pas empêché certains auteurs de réussir à mener des méta-analyses complètes de jeux de données utilisant des puces à ADN.

Par exemple, Gur-Dedeoglu et ses collègues [199] ont proposé une méthode de méta-analyse ayant mis en évidence un ensemble de gènes dont l'expression permet de différencier un tissu mammaire sain et des carcinomes canalaires ou lobulaires invasifs. Pour cela, ils ont sélectionné deux jeux de données similaires sur lesquels ils ont appliqué une méthode de ré-échantillonnage afin de tester différentes hypothèses. Cette technique présente l'avantage de pouvoir mettre en évidence certains changements trop faibles pour être détectés dans un seul jeu de données. Cependant, le ré-échantillonnage empêche de considérer un grand nombre de jeux de données, à moins que de disposer d'une très importante capacité de calcul. De plus, cette méthode rend l'utilisation de plusieurs types de puces à ADN impossible.

Lors d'une méta-analyse comparative, les résultats d'une étude peuvent être renforcés par les résultats d'autres études ou des hypothèses peuvent se voir écartées car elles ne sont validées que par trop peu de jeux de données. Cette approche a donc le potentiel de permettre de nouvelles découvertes. Cependant, différents problèmes empêchent la comparaison de listes de gènes d'être efficace. En effet, en plus d'être souvent très longues, elles utilisent en général des nomenclatures différentes. De plus, les listes de gènes exprimés de manière différentielle sont souvent publiées sous formes de graphiques ou de tableaux inaccessibles aux outils de recherche. Enfin, malgré que les données brutes générées par les puces à ADN soient de plus en plus fréquemment disponibles publiquement, les analyses des résultats sont, elles, rarement disponibles dans les bases de données telles que GEO et ArrayExpress. Toutefois, des chercheurs sont parvenus à contourner ces obstacles pour fournir des méta-analyses comparatives de qualité.

Citons, par exemple, l'étude de Wennmalm et de ses collègues [214] qui, en comparant les résultats de sept jeux de données utilisant des puces à ADN, ont pu mettre en évidence un groupe de gènes impliqués dans le vieillissement chez plusieurs espèces : la souris, le rat et l'être humain. De plus, ils ont montré qu'il existait une forte similarité entre un ensemble de gènes impliqués dans la sénéscence cellulaire et un ensemble de gènes impliqués dans le vieillissement chez la souris mais pas chez l'homme. À partir de 32 jeux de données impliquant 77 interactions différentes entre pathogènes et hôtes, Jenner et Young [215] ont,

eux, défini un ensemble de gènes systématiquement transcrits chez l'hôte lors d'une infection. Ceux-ci étaient, pour la plupart, liés à la voie de signalisation des toll-like recepteurs. Nous voyons donc que les méta-analyses peuvent, non seulement prendre en compte un grand nombre de jeux de données, mais en plus, ceux-ci peuvent avoir été générés avec des modèles de puce à ADN différents. Cependant, il faut garder à l'esprit que ce type de méta-analyse regroupe des études dont les résultats ont été obtenus par des méthodes qui peuvent être différentes. Sachant que les résultats générés par différentes méthodes peuvent, eux-mêmes, être très différents, les méta-analyses comparatives peuvent comporter un biais important à ce niveau.

Outre les méta-analyses complètes et comparatives, il est nécessaire de signaler qu'il existe aussi une troisième catégorie intermédiaire. Dans ce type d'approche, les jeux de données ne sont pas rassemblés pour une analyse unique comme c'est le cas pour les méta-analyses complètes, mais ils sont tout de même ré-analysés individuellement, contrairement aux méta-analyses comparatives, afin que les résultats soient tous générés par les mêmes méthodes.

C'est ainsi que Ma et Huang ont ré-analysé quatre jeux de données relatifs aux adénocarcinomes pancréatiques avec la méthode appelée « Threshold Gradient Descent Regularization » (TGDR) [200]. En intégrant les résultats de ces analyses à une méthode qu'ils ont appelé « Meta Threshold Gradient Descent Regularization » (MTGDR), ils ont été capables d'identifier un ensemble de gènes responsables du développement de cette maladie. On voit ici que relativement peu de jeux de données ont été intégrés dans la méta-analyse. Cependant, elle présentait l'avantage de prendre en considération des jeux de données utilisant trois technologies différentes.

Avec l'outil « Gene Expression MetaSignatures » (GEMS), Ochsner et ses collègues [201] ont sélectionné dix jeux de données relatifs au traitement de cellules MCF-7 avec du 17beta-estradiol. Ils ont combiné les valeurs, déjà prétraitées avec RMA ou GCRMA, des 13.000 gènes représentés à la fois sur les deux modèles de puces à ADN utilisés pour mettre en évidence des groupes de gènes impliqués dans la réponse cellulaire à court terme (après 3 ou 4 heures) et à long terme (après 24 heures) au 17beta-estradiol. Bien que cette application de GEMS prenne en considération un nombre relativement important de jeux de données, le nombre de modèles de puce à ADN impliqués ne peut pas être trop élevé sous peine de trop restreindre le nombre de gènes analysés. En effet, GEMS ne prend en compte que les gènes qui sont représentés sur tous les modèles de puce à ADN impliqués dans la méta-analyse.

Comme le nombre de gènes communs représentés sur tous les modèles de puce à ADN peut fortement diminuer quand de nombreux modèles sont impliqués, cette situation doit être évitée lors de l'utilisation de GEMS.

Comme la méta-analyse que nous proposons est composée de plusieurs approches différentes, elle appartient à plusieurs catégories. Premièrement, les approches nommées « intersections » et « intersections d'unions » appartiennent à la catégorie intermédiaire entre les méta-analyses complètes et comparatives puisqu'elles reposent sur les résultats des ré-analyses individuelles des jeux de données. Comme ces ré-analyses ont été réalisées avec une panoplie d'outils identiques, ceci nous permet d'obtenir une cohérence plus grande entre les différents résultats obtenus par les intersections et les intersections d'unions. Ensuite, l'approche que nous avons appelée les « méta-analyses » peut être classées dans la catégorie des méta-analyses complètes. En effet, avec cette approche, nous assemblons plusieurs jeux de données en un que nous prétraitons et traitons de manière classique afin d'augmenter la puissance statistique de l'analyse.

La combinaison de ces trois approches met en avant plusieurs avantages : d'abord, le nombre de jeux de données impliqués peut être très grand. De plus, ceux-ci peuvent être générés avec différents modèles de puce à ADN. Ensuite, les outils utilisés sont toujours les mêmes. Enfin, l'élément totalement novateur dans notre méthodologie est la possibilité de prendre en compte, au sein d'une même méta-analyse, des jeux de données relatifs à plusieurs processus biologiques différents.

Appliquée aux jeux de données que nous avons sélectionnés, notre méta-analyse a permis de mettre en évidence une implication possible de deux voies de signalisation : « toll-like receptor signalling pathway » et « Fc gamma R-mediated phagocytosis ». L'implication de ces voies dans le cancer a rarement été mise en évidence. Cela fait longtemps que les agonistes des toll-like recepteurs (TLR) sont utilisés comme immunoadjuvants dans la thérapie contre le cancer. Cependant, un nombre croissant d'études commencent à rapporter un rôle éventuel des toll-like recepteurs dans le développement cancéreux.

Récemment, French et ses collègues ont publié une revue présentant deux exemples de carcinomes hépatiques développés suite à l'activation des voies de signalisation des TLRs 2 et 4 [216]. En effet, dans le premier exemple, l'apport chronique d'éthanol ou de lipopolysaccharides à des souris a entraîné l'activation de la voie du TLR 4. Dans le second exemple, c'est l'apport de diethyl 1,4-dehydro-2,3,6-trimethyl-3,5-pyridine decarboxylate (DDC) dans la nourriture des souris qui a permis d'activer les voies des TLRs 2 et 4. Dans les

deux cas, l'activation des TLRs a conduit à la prolifération et la transformation de cellules souches en cellules hépatiques cancéreuses. De plus, l'inactivation des TLRs ou de leur action permettait d'éviter cette transformation.

Zheng et son équipe ont également montré que les hommes présentant une variation particulière dans la séquence du gène codant pour le TLR 4 avaient une augmentation de 26% du risque de développer un cancer de la prostate [217]. De plus, leur étude indiquait une relation statistiquement significative entre la présence d'un SNP particulier dans les gènes des TLRs 4, 6 ou 10 et un risque élevé de développer un cancer de la prostate.

Chang et ses collègues ont montré que, lors d'une infection par *Helicobacter pylori*, les voies de signalisation en aval des TLRs 2 et 9 étaient activées, ce qui entraînait l'expression de COX-2 qui pouvait contribuer potentiellement à la progression d'un cancer gastrique [218].

Les indications d'une implication des TLRs et des voies de signalisation qu'ils déclenchent dans le développement tumoral s'accumulent donc. Ce n'est d'ailleurs pas si étonnant. En effet, il est de plus en plus accepté que les cellules cancéreuses ont développé des mécanismes pour échapper au système immunitaire, voire même en tirer parti. L'observation de l'équipe de Huang [219], selon laquelle l'activation de la voie du TLR 4 par des lipopolysaccharides dans des cellules cancéreuses induisait la synthèse de facteurs tels que les interleukines 6 et 12 ou iNOS leur permettant ainsi de résister aux lymphocytes T, ne fait que renforcer cette hypothèse. Les résultats que nous présentons ici sont un élément de plus, et ils vont même plus loin puisque nous proposons que la voie de signalisation des toll-like recepteurs pourrait être impliquée dans le développement métastatique.

L'implication de la voie de signalisation « Fc Gamma R-mediated phagocytosis » dans le cancer ou le phénotype métastatique a encore été plus rarement rapportée que la voie des TLRs. Pourtant, il s'agit d'un processus où le cytosquelette joue un rôle important. Sachant qu'il y a une réorganisation du cytosquelette lorsque les cellules cancéreuses migrent, il n'est pas étonnant de voir des gènes impliqués dans la phagocytose être sollicités lors de la cascade métastatique. En effet, la voie de signalisation « Fc gamma R-mediated phagocytosis » est un processus mis en place par les macrophages, les neutrophiles et les monocytes pour éliminer une menace pathogène. Après la reconnaissance extracellulaire d'une molécule venant d'un organisme pathogène par un récepteur Fc gamma, un signal intracellulaire induit la formation d'un phagosome qui fusionne ensuite avec des lysosomes. Les protéases lysosomales permettent alors de digérer le pathogène. La formation du phagosome exige des changements

dans la structure et la dynamique du cytosquelette similaires à ceux que l'on peut observer dans des cellules cancéreuses qui migrent d'une tumeur primaire vers un site secondaire [220]. En effet, pour migrer, les cellules cancéreuses doivent s'allonger, émettre des pseudopodes ou encore se contracter. Toutes ces étapes demandent, notamment, la dépolymérisation/repolymérisation de l'actine et la contraction des filaments qu'elle forme par l'action de la myosine II. Ces mécanismes moléculaires sont similaires à ceux mis en place lors de la phagocytose [83]. Notre étude fait partie des premières à suggérer que le processus de migration des cellules cancéreuses peut s'opérer par la régulation des gènes également impliqués dans la phagocytose.

Ces différents éléments indiquant un rôle potentiel important des voies de signalisation « toll like receptor signalling pathway » et « Fc Gamma R-mediated phagocytosis » dans le développement métastatique nous ont conduit à les investiguer plus en profondeur. Nous avons pu montrer par des approches expérimentales que l'invalidation de PAK1 ou de CFL2, de la voie de signalisation « Fc Gamma R-mediated phagocytosis », entrainait une perte des capacités migratoires des cellules MDA-MB-231, qui sont des cellules à haut potentiel métastatique.

Bien que cette étude suggère une implication des voies de signalisation « Fc Gamma R-mediated phagocytosis » et « toll like receptor signalling pathway » dans le cancer et qu'elle soit la première à montrer une régulation possible de l'expression des gènes de la phagocytose dans le développement métastatique, il reste une large gamme de perspectives qu'il serait intéressant d'approfondir.

D'abord au niveau de la méthodologie de méta-analyse de données issues de puces à ADN, nous recommandons, pour toute utilisation future, la mise à jour des CDFs, l'utilisation du shrinkage t test, le développement d'une formule statistique pour le calcul du rang seuil des intersections d'unions et l'utilisation d'une version récente du programme R. De plus, il serait intéressant de développer une interface automatisant tout le processus de la méta-analyse. En effet, nous avons montré que la méthodologie que nous proposons est un outil puissant. Cependant, bien qu'elles soient simples, les approches que nous avons décrites demandent quelques compétences bioinformatiques pour être implémentées, les rendant, dès lors, relativement inaccessibles aux autres biologistes qui voudraient générer de nouvelles hypothèses à partir de puces à ADN. Ensuite, notre méthodologie de méta-analyse serait rendue encore plus puissante si elle pouvait intégrer des jeux de données issus d'autres types

de puce à ADN que celles développées par Affymetrix. De plus, cela nous permettrait d'accéder à une masse d'informations encore plus grande.

Durant ce projet, nous avons proposé que les 74 gènes d'intérêt (165 – 91) qui n'ont pas encore été décrits dans la littérature comme étant impliqués dans les cancers, le développement métastatique ou la réponse à l'hypoxie constituaient de bons candidats à la validation expérimentale. Cependant, afin de s'assurer que ces 74 gènes ne constituent pas le bruit de fond de la méta-analyse, nous pourrions permuter certaines puces à ADN d'une condition à l'autre au sein des jeux de données qui ont servi à la méta-analyse. De cette manière, au sein d'un jeu de données, la première condition contiendrait des puces à ADN où de l'ARN de cellules métastatiques ou non métastatiques a été hybridé tout comme la seconde condition (la même permutation de conditions devrait être opérée pour les jeux de données comparant des cellules incubées en hypoxie à des cellules incubées en normoxie). Ainsi, l'hypothèse nulle selon laquelle ni le phénotype métastatique, ni la réponse à l'hypoxie n'ont d'effet serait assurée. En effet, pour chaque jeu de données l'analyse comparerait l'ensemble des puces d'une condition à l'ensemble des puces de l'autre condition. Si les deux conditions comportent toutes deux des puces à ADN où de l'ARN de cellules métastatiques et non métastatiques (ou incubées en normoxie et en hypoxie) a été hybridé, aucun effet dû aux conditions ne devrait plus être observé. De cette façon, si la méta-analyse que nous proposons n'est pas sujette au bruit de fond, les résultats devraient être nuls et aucun gène ne devrait être sélectionné par la méta-analyse. Par contre, un nombre de gènes proche de 74 devrait être sélectionné par la méta-analyse si le bruit de fond est trop important.

Enfin, la robustesse de la méta-analyse pourrait être testée en diminuant le nombre de jeux de données pour voir si les résultats en seraient plus ou moins affectés. Dans l'article présentant la première version de la méta-analyse que nous proposons [207], nous avions déjà réalisé ce type de test. En effet, en ne prenant en compte que les jeux de données comparant des cellules métastatiques aux cellules de mélanomes dont elles se sont détachées, la méta-analyse ne sélectionnait que 7 gènes. De même, en ne prenant en compte que les jeux de données comparant des cellules métastatiques aux cellules de tumeur primaire de prostate dont elles se sont détachées, la méta-analyse ne sélectionnait que 17 gènes. La méta-analyse permet donc la sélection d'un plus grand nombre de gènes lorsqu'elle est enrichie en jeux de données. Cependant, il serait intéressant de mener les mêmes tests en éliminant des jeux de données sans tenir compte du type de tissu qu'ils étudient. De plus, il serait aussi intéressant de mener les mêmes tests en éliminant les jeux de données comparant des cellules incubées en

hypoxie et en normoxie afin de voir quel est l'apport réel de ces jeux de données à la méta-analyse.

Au niveau des expériences de validation expérimentale que nous avons effectuées, les différents résultats présentés indiquent une implication possible de PAK1 et de CFL2 dans le potentiel migratoire des cellules cancéreuses. Cependant, des expériences *in vivo* constituent l'étape suivante dans la validation de leur rôle dans le processus métastatique proprement dit. Pour cela, les cellules contrôles ou transfectées de façon stable avec un siRNA devraient être injectées dans la veine de la queue de souris. Six semaines après, les souris seraient sacrifiées et les poumons seraient prélevés, fixés et analysés par coupe histologique pour la présence de métastases, c'est-à-dire de cellules MDA-MB-231. Le rôle de PAK1 et de CFL2 pourrait ainsi être évalué dans l'établissement des métastases *in vivo*. De plus, d'autres expériences *in vitro* pourraient être réalisées. En effet, nous avons montré le rôle de PAK1 et de CFL2 en invalidant leur expression dans des cellules à haut potentiel métastatique, mais l'expérience inverse pourrait également être menée en surexprimant ces gènes dans une lignée à bas potentiel métastatique (par exemple, les cellules MCF-7). Ensuite, les résultats concernant la migration cellulaire que nous avons obtenus pourraient être validés par l'utilisation d'autres tests de migration, par exemple en utilisant des chambres de Boyden. De plus, les « scratch assays » que nous avons réalisés ne prenaient en compte qu'un milieu cellulaire à deux dimensions. En effet, les grattes et l'observation de leur résorption se sont faites lorsqu'une monocouche de cellules était formée. Pour simuler un milieu cellulaire en trois dimensions, nous pourrions utiliser des inserts de Matrigel lors des tests de migration cellulaire, que ce soit par « scratch assays » ou en chambre de Boyden. Ainsi, nous serions dans des conditions plus proches de la réalité biologique.

Finalement, une étude rétrospective sur PAK1 et CFL2 pourrait être réalisée sur des patientes atteintes d'un cancer du sein. Il s'agirait de déterminer l'expression de l'ARNm de PAK1 et de CFL2 à partir des données de puces à ADN déjà disponibles et si cela s'avère possible, l'expression de la protéine sur coupes histologiques et de suivre la rémission sans métastase et la survie des patientes séparées en deux groupes en fonction de ces données d'expression, sur des courbes de Kaplan-Meyer. Il serait alors possible de voir si une expression élevée de ces gènes est associée à une survie moindre des patientes.

Pour conclure, nous avons mis au point une méthode de méta-analyse de puces à ADN qui permet l'analyse simultanée de plusieurs processus biologiques, ce qui en fait un outil original et novateur. Appliqué à des jeux de données traitant de métastases et de réponse à

l'hypoxie, cet outil nous a permis de mettre en évidence 165 gènes d'intérêt. Parmi ceux-ci, un grand nombre étaient déjà connus pour être impliqués dans le développement cancéreux et/ou dans la réponse à l'hypoxie, montrant la puissance de la méthode.

Parmi les voies de signalisation dans lesquelles les 165 gènes d'intérêt étaient impliqués, plusieurs étaient liées à la reconnaissance de pathogènes et à la phagocytose. En particulier, les voies « Fc Gamma R-mediated phagocytosis » et « toll like receptor signalling pathway » ont rarement été rapportées dans la littérature comme participant à la migration des cellules cancéreuses. Des expériences que nous avons réalisées sur cellules ont montré une possible implication de deux gènes (PAK1 et CFL2) de la voie « Fc Gamma R-mediated phagocytosis » dans le potentiel métastatique des cellules cancéreuses. En effet, lors de tests de migration, l'invalidation de ces deux gènes, par transfection de siRNA dans des cellules cancéreuses hautement métastatiques, a causé une importante diminution de la capacité migratoire des cellules.

Outre l'indication novatrice que les voies de signalisation « Fc Gamma R-mediated phagocytosis » et « toll like receptor signalling pathway » sont impliqués dans le processus métastatique, ces résultats positionnent, pour la première fois, PAK1 et CFL2 comme de bonnes cibles pour de futures thérapies contre cette maladie. Bien sûr, d'autres expériences sont encore nécessaires pour valider le rôle de ces gènes dans le processus métastatique.

V. REFERENCES

1. Vogelstein B, Kinzler KW: **The multistep nature of cancer**. *Trends Genet* 1993, **9**(4):138-141.

2. Tiainen M, Pajalunga D, Ferrantelli F, Soddu S, Salvatori G, Sacchi A, Crescenzi M: **Terminally differentiated skeletal myotubes are not confined to G0 but can enter G1 upon growth factor stimulation**. *Cell Growth Differ* 1996, **7**(8):1039-1050.

3. Elledge SJ: **Cell cycle checkpoints: preventing an identity crisis**. *Science* 1996, **274**(5293):1664-1672.

4. Giuriato S, Felsher DW: **How cancers escape their oncogene habit**. *Cell Cycle* 2003, **2**(4):329-332.

5. Kaufmann WK, Kaufman DG: **Cell cycle control, DNA repair and initiation of carcinogenesis**. *Faseb J* 1993, **7**(12):1188-1191.

6. Ozaki T, Nakagawara A: **p53: the attractive tumor suppressor in the cancer research field**. *J Biomed Biotechnol* 2011, **2011**:603925.

7. Kinzler KW, Vogelstein B: **Cancer-susceptibility genes. Gatekeepers and caretakers**. *Nature* 1997, **386**(6627):761, 763.

8. Jarvinen HJ, Aarnio M: **Surveillance on mutation carriers of DNA mismatch repair genes**. *Ann Chir Gynaecol* 2000, **89**(3):207-210.

9. Sulic S, Panic L, Dikic I, Volarevic S: **Deregulation of cell growth and malignant transformation**. *Croat Med J* 2005, **46**(4):622-638.

10. Negrini S, Gorgoulis VG, Halazonetis TD: **Genomic instability--an evolving hallmark of cancer**. *Nat Rev Mol Cell Biol* 2010, **11**(3):220-228.

11. Dvorak HF: **Tumors: wounds that do not heal. Similarities between tumor stroma generation and wound healing**. *N Engl J Med* 1986, **315**(26):1650-1659.

12. Pages F, Galon J, Dieu-Nosjean MC, Tartour E, Sautes-Fridman C, Fridman WH: **Immune infiltration in human tumors: a prognostic factor that should not be ignored**. *Oncogene* 2010, **29**(8):1093-1102.

13. DeNardo DG, Andreu P, Coussens LM: **Interactions between lymphocytes and myeloid cells regulate pro- versus anti-tumor immunity**. *Cancer Metastasis Rev* 2010, **29**(2):309-316.

14. Grivennikov SI, Greten FR, Karin M: **Immunity, inflammation, and cancer**. *Cell* 2010, **140**(6):883-899.

15. Schmitz S, Machiels JP: **Molecular biology of squamous cell carcinoma of the head and neck: relevance and therapeutic implications**. *Expert Rev Anticancer Ther* 2010, **10**(9):1471-1484.

16. Yu Z, Ren P, Zhang X, Zhang T, Ma B: **Therapeutic potential of dendritic cell vaccines in sarcoma of the extremities**. *Expert Rev Anticancer Ther* 2009, **9**(8):1065-1071.

17. Drexler HG, Macleod RA: **History of leukemia-lymphoma cell lines**. *Hum Cell* 2010, **23**(3):75-82.

18. Jessberger R: **New insights into germ cell tumor formation**. *Horm Metab Res* 2008, **40**(5):342-346.

19. Romeo C, Impellizzeri P, Grosso M, Vitarelli E, Gentile C: **Pleuropulmonary blastoma: long-term survival and literature review**. *Med Pediatr Oncol* 1999, **33**(4):372-376.

20. Crepin M, Dieu MC, Lejeune S, Escande F, Boidin D, Porchet N, Morin G, Manouvrier S, Mathieu M, Buisine MP: **Evidence of constitutional MLH1 epimutation associated to transgenerational inheritance of cancer susceptibility**. *Hum Mutat* 2011.

21. Blanchetot C, Boonstra J: **The ROS-NOX connection in cancer and angiogenesis**. *Crit Rev Eukaryot Gene Expr* 2008, **18**(1):35-45.

22. Cleaver JE, Crowley E: **UV damage, DNA repair and skin carcinogenesis**. *Front Biosci* 2002, **7**:d1024-1043.

23. Seitz HK, Becker P: **Alcohol metabolism and cancer risk**. *Alcohol Res Health* 2007, **30**(1):38-41, 44-37.

24. Sasco AJ, Secretan MB, Straif K: **Tobacco smoking and cancer: a brief review of recent epidemiological evidence**. *Lung Cancer* 2004, **45 Suppl 2**:S3-9.

25. Rezazadeh A, Laber DA, Ghim SJ, Jenson AB, Kloecker G: **The role of human papilloma virus in lung cancer: a review of the evidence**. *Am J Med Sci* 2009, **338**(1):64-67.

26. Hanahan D, Weinberg RA: **The hallmarks of cancer**. *Cell* 2000, **100**(1):57-70.

27. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation**. *Cell* 2011, **144**(5):646-674.

28. Yano S, Kondo K, Yamaguchi M, Richmond G, Hutchison M, Wakeling A, Averbuch S, Wadsworth P: **Distribution and function of EGFR in human tissue and the effect of EGFR tyrosine kinase inhibition**. *Anticancer Res* 2003, **23**(5A):3639-3650.

29. Cheng N, Chytil A, Shyr Y, Joly A, Moses HL: **Transforming growth factor-beta signaling-deficient fibroblasts enhance hepatocyte growth factor signaling in mammary carcinoma cells to promote scattering and invasion**. *Mol Cancer Res* 2008, **6**(10):1521-1533.

30. Evan GI, d'Adda di Fagagna F: **Cellular senescence: hot or what?** *Curr Opin Genet Dev* 2009, **19**(1):25-31.

31. Collado M, Serrano M: **Senescence in tumours: evidence from mice and humans**. *Nat Rev Cancer* 2010, **10**(1):51-57.

32. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL: **Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene**. *Science* 1987, **235**(4785):177-182.

33. Yarden Y, Ullrich A: **Growth factor receptor tyrosine kinases**. *Annu Rev Biochem* 1988, **57**:443-478.

34. Fedi P, Tronick SR, Aaronson SA: **Growth factors**. Baltimore: Williams and Wilkins; 1997.

35. Giancotti FG, Ruoslahti E: **Integrin signaling**. *Science* 1999, **285**(5430):1028-1032.

36. Jiang BH, Liu LZ: **PI3K/PTEN signaling in angiogenesis and tumorigenesis**. *Adv Cancer Res* 2009, **102**:19-65.

37. Burkhart DL, Sage J: **Cellular mechanisms of tumour suppression by the retinoblastoma gene**. *Nat Rev Cancer* 2008, **8**(9):671-682.

38. Lipinski MM, Jacks T: **The retinoblastoma gene family in differentiation and development**. *Oncogene* 1999, **18**(55):7873-7882.

39. Ghebranious N, Donehower LA: **Mouse models in tumor suppression**. *Oncogene* 1998, **17**(25):3385-3400.

40. Moses HL, Yang EY, Pietenpol JA: **TGF-beta stimulation and inhibition of cell proliferation: new mechanistic insights**. *Cell* 1990, **63**(2):245-247.

41. Zuo L, Weger J, Yang Q, Goldstein AM, Tucker MA, Walker GJ, Hayward N, Dracopoli NC: **Germline mutations in the p16INK4a binding domain of CDK4 in familial melanoma**. *Nat Genet* 1996, **12**(1):97-99.

42. Datto MB, Hu PP, Kowalik TF, Yingling J, Wang XF: **The viral oncoprotein E1A blocks transforming growth factor beta-mediated induction of p21/WAF1/Cip1 and p15/INK4B**. *Mol Cell Biol* 1997, **17**(4):2030-2037.

43. Dyson N, Howley PM, Munger K, Harlow E: **The human papilloma virus-16 E7 oncoprotein is able to bind to the retinoblastoma gene product**. *Science* 1989, **243**(4893):934-937.

44. Foley KP, Eisenman RN: **Two MAD tails: what the recent knockouts of Mad1 and Mxi1 tell us about the MYC/MAX/MAD network**. *Biochim Biophys Acta* 1999, **1423**(3):M37-47.

45. Kinzler KW, Vogelstein B: **Lessons from hereditary colorectal cancer**. *Cell* 1996, **87**(2):159-170.

46. Curto M, Cole BK, Lallemand D, Liu CH, McClatchey AI: **Contact-dependent inhibition of EGFR signaling by Nf2/Merlin**. *J Cell Biol* 2007, **177**(5):893-903.

47. Shaw RJ: **Tumor suppression by LKB1: SIK-ness prevents metastasis**. *Sci Signal* 2009, **2**(86):pe55.

48. Lowe SW, Cepero E, Evan G: **Intrinsic tumour suppression**. *Nature* 2004, **432**(7015):307-315.

49. Kerr JF, Wyllie AH, Currie AR: **Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics**. *Br J Cancer* 1972, **26**(4):239-257.

50. Adams JM, Cory S: **The Bcl-2 apoptotic switch in cancer development and therapy**. *Oncogene* 2007, **26**(9):1324-1337.

51. Junttila MR, Evan GI: **p53--a Jack of all trades but master of none**. *Nat Rev Cancer* 2009, **9**(11):821-829.

52. Thornberry NA, Lazebnik Y: **Caspases: enemies within**. *Science* 1998, **281**(5381):1312-1316.

53. Harris CC: **p53 tumor suppressor gene: from the basic research laboratory to the clinic--an abridged historical perspective**. *Carcinogenesis* 1996, **17**(6):1187-1198.

54. Cantley LC, Neel BG: **New insights into tumor suppression: PTEN suppresses tumor formation by restraining the phosphoinositide 3-kinase/AKT pathway**. *Proc Natl Acad Sci U S A* 1999, **96**(8):4240-4245.

55. Evan G, Littlewood T: **A matter of life and cell death**. *Science* 1998, **281**(5381):1317-1322.

56. Amaral JD, Xavier JM, Steer CJ, Rodrigues CM: **The role of p53 in apoptosis**. *Discov Med* 2010, **9**(45):145-152.

57. Hayflick L: **Mortality and immortality at the cellular level. A review**. *Biochemistry (Mosc)* 1997, **62**(11):1180-1190.

58. Kawai T, Hiroi S, Nakanishi K, Meeker AK: **Telomere length and telomerase expression in atypical adenomatous hyperplasia and small bronchioloalveolar carcinoma of the lung**. *Am J Clin Pathol* 2007, **127**(2):254-262.

59. Blasco MA: **Telomeres and human disease: ageing, cancer and beyond**. *Nat Rev Genet* 2005, **6**(8):611-622.

60. Counter CM, Avilion AA, LeFeuvre CE, Stewart NG, Greider CW, Harley CB, Bacchetti S: **Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activity**. *Embo J* 1992, **11**(5):1921-1929.

61. Wright WE, Pereira-Smith OM, Shay JW: **Reversible cellular senescence: implications for immortalization of normal human diploid fibroblasts**. *Mol Cell Biol* 1989, **9**(7):3088-3092.

62. Bryan TM, Cech TR: **Telomerase and the maintenance of chromosome ends**. *Curr Opin Cell Biol* 1999, **11**(3):318-324.

63. Veikkola T, Alitalo K: **VEGFs, receptors and angiogenesis**. *Semin Cancer Biol* 1999, **9**(3):211-220.

64. Kazerounian S, Yee KO, Lawler J: **Thrombospondins in cancer**. *Cell Mol Life Sci* 2008, **65**(5):700-712.

65. Baeriswyl V, Christofori G: **The angiogenic switch in carcinogenesis**. *Semin Cancer Biol* 2009, **19**(5):329-337.

66. Mac Gabhann F, Popel AS: **Systems biology of vascular endothelial growth factors**. *Microcirculation* 2008, **15**(8):715-738.

67. Dameron KM, Volpert OV, Tainsky MA, Bouck N: **Control of angiogenesis in fibroblasts by p53 regulation of thrombospondin-1**. *Science* 1994, **265**(5178):1582-1584.

68. Su F, Pascal LE, Xiao W, Wang Z: **Tumor suppressor U19/EAF2 regulates thrombospondin-1 expression via p53**. *Oncogene* 2010, **29**(3):421-431.

69. Nagy JA, Chang SH, Shih SC, Dvorak AM, Dvorak HF: **Heterogeneity of the tumor vasculature**. *Semin Thromb Hemost* 2010, **36**(3):321-331.

70. Hayes AJ, Li LY, Lippman ME: **Science, medicine, and the future. Antivascular therapy: a new approach to cancer treatment**. *Bmj* 1999, **318**(7187):853-856.

71. McDonald DM, Choyke PL: **Imaging of angiogenesis: from microscope to clinic**. *Nat Med* 2003, **9**(6):713-725.

72. Fidler IJ: **The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited**. *Nat Rev Cancer* 2003, **3**(6):453-458.

73. Warburg O: **On respiratory impairment in cancer cells**. *Science* 1956, **124**(3215):269-270.

74. Hsu PP, Sabatini DM: **Cancer cell metabolism: Warburg and beyond**. *Cell* 2008, **134**(5):703-707.

75. Semenza GL: **HIF-1: upstream and downstream of cancer metabolism**. *Curr Opin Genet Dev* 2010, **20**(1):51-56.

76. Potter VR: **The biochemical approach to the cancer problem**. *Fed Proc* 1958, **17**(2):691-697.

77. Vander Heiden MG, Cantley LC, Thompson CB: **Understanding the Warburg effect: the metabolic requirements of cell proliferation**. *Science* 2009, **324**(5930):1029-1033.

78. Feron O: **Pyruvate into lactate and back: from the Warburg effect to symbiotic energy fuel exchange in cancer cells**. *Radiother Oncol* 2009, **92**(3):329-333.

79. Teng MW, Swann JB, Koebel CM, Schreiber RD, Smyth MJ: **Immune-mediated dormancy: an equilibrium with cancer**. *J Leukoc Biol* 2008, **84**(4):988-993.

80. Nelson BH: **The impact of T-cell immunity on ovarian cancer outcomes**. *Immunol Rev* 2008, **222**:101-116.

81. Shields JD, Kourtis IC, Tomei AA, Roberts JM, Swartz MA: **Induction of lymphoidlike stroma and immune escape by tumors that express the chemokine CCL21**. *Science* 2010, **328**(5979):749-752.

82. Mougiakakos D, Choudhury A, Lladser A, Kiessling R, Johansson CC: **Regulatory T cells in cancer**. *Adv Cancer Res* 2010, **107**:57-117.

83. Friedl P, Wolf K: **Tumour-cell invasion and migration: diversity and escape mechanisms**. *Nat Rev Cancer* 2003, **3**(5):362-374.

84. Pantel K, Brakenhoff RH: **Dissecting the metastatic cascade**. *Nat Rev Cancer* 2004, **4**(6):448-456.

85. Sullivan R, Graham CH: **Hypoxia-driven selection of the metastatic phenotype**. *Cancer Metastasis Rev* 2007, **26**(2):319-331.

86. Bienz M: **beta-Catenin: a pivot between cell adhesion and Wnt signalling**. *Curr Biol* 2005, **15**(2):R64-67.

87. Graff JR, Gabrielson E, Fujii H, Baylin SB, Herman JG: **Methylation patterns of the E-cadherin 5' CpG island are unstable and reflect the dynamic, heterogeneous loss of E-cadherin expression during metastatic progression**. *J Biol Chem* 2000, **275**(4):2727-2732.

88. Krishnamachary B, Zagzag D, Nagasawa H, Rainey K, Okuyama H, Baek JH, Semenza GL: **Hypoxia-inducible factor-1-dependent repression of E-cadherin in von Hippel-Lindau tumor suppressor-null renal cell carcinoma mediated by TCF3, ZFHX1A, and ZFHX1B**. *Cancer Res* 2006, **66**(5):2725-2731.

89. Imai T, Horiuchi A, Wang C, Oka K, Ohira S, Nikaido T, Konishi I: **Hypoxia attenuates the expression of E-cadherin via up-regulation of SNAIL in ovarian carcinoma cells**. *Am J Pathol* 2003, **163**(4):1437-1447.

90. Bates RC, Mercurio AM: **The epithelial-mesenchymal transition (EMT) and colorectal cancer progression**. *Cancer Biol Ther* 2005, **4**(4):365-370.

91. Kuphal S, Poser I, Jobin C, Hellerbrand C, Bosserhoff AK: **Loss of E-cadherin leads to upregulation of NFkappaB activity in malignant melanoma**. *Oncogene* 2004, **23**(52):8509-8519.

92. Sacco PA, McGranahan TM, Wheelock MJ, Johnson KR: **Identification of plakoglobin domains required for association with N-cadherin and alpha-catenin**. *J Biol Chem* 1995, **270**(34):20201-20206.

93. Li G, Satyamoorthy K, Herlyn M: **N-cadherin-mediated intercellular interactions promote survival and migration of melanoma cells**. *Cancer Res* 2001, **61**(9):3819-3825.

94. Atouf F, Park CH, Pechhold K, Ta M, Choi Y, Lumelsky NL: **No evidence for mouse pancreatic beta-cell epithelial-mesenchymal transition in vitro**. *Diabetes* 2007, **56**(3):699-702.

95. Chase LG, Ulloa-Montoya F, Kidder BL, Verfaillie CM: **Islet-derived fibroblast-like cells are not derived via epithelial-mesenchymal transition from Pdx-1 or insulin-positive cells**. *Diabetes* 2007, **56**(1):3-7.

96. Morton RA, Geras-Raaka E, Wilson LM, Raaka BM, Gershengorn MC: **Endocrine precursor cells from mouse islets are not generated by epithelial-to-mesenchymal transition of mature beta cells**. *Mol Cell Endocrinol* 2007, **270**(1-2):87-93.

97. Weinberg N, Ouziel-Yahalom L, Knoller S, Efrat S, Dor Y: **Lineage tracing evidence for in vitro dedifferentiation but rare proliferation of mouse pancreatic beta-cells**. *Diabetes* 2007, **56**(5):1299-1304.

98. Joyce JA, Baruch A, Chehade K, Meyer-Morse N, Giraudo E, Tsai FY, Greenbaum DC, Hager JH, Bogyo M, Hanahan D: **Cathepsin cysteine proteases are effectors of invasive growth and angiogenesis during multistage tumorigenesis**. *Cancer Cell* 2004, **5**(5):443-453.

99. Kos J, Lah TT: **Cysteine proteinases and their endogenous inhibitors: target proteins for prognosis, diagnosis and therapy in cancer (review)**. *Oncol Rep* 1998, **5**(6):1349-1361.

100. Nomura T, Katunuma N: **Involvement of cathepsins in the invasion, metastasis and proliferation of cancer cells**. *J Med Invest* 2005, **52**(1-2):1-9.

101. Westermarck J, Kahari VM: **Regulation of matrix metalloproteinase expression in tumor invasion**. *Faseb J* 1999, **13**(8):781-792.

102. Sternlicht MD, Werb Z: **How matrix metalloproteinases regulate cell behavior**. *Annu Rev Cell Dev Biol* 2001, **17**:463-516.

103. Noel A, Jost M, Maquoi E: **Matrix metalloproteinases at cancer tumor-host interface**. *Semin Cell Dev Biol* 2008, **19**(1):52-60.

104. Liotta LA, Goldfarb RH, Brundage R, Siegal GP, Terranova V, Garbisa S: **Effect of plasminogen activator (urokinase), plasmin, and thrombin on glycoprotein and collagenous components of basement membrane**. *Cancer Res* 1981, **41**(11 Pt 1):4629-4636.

105. Vempati P, Mac Gabhann F, Popel AS: **Quantifying the proteolytic release of extracellular matrix-sequestered VEGF with a computational model**. *PLoS One* 2010, **5**(7):e11860.

106. Moroy G, Bourguet E, Decarme M, Sapi J, Alix AJ, Hornebeck W, Lorimier S: **Inhibition of human leukocyte elastase, plasmin and matrix metalloproteinases by oleic acid and oleoyl-galardin derivative(s)**. *Biochem Pharmacol* 2011, **81**(5):626-635.

107. Yebra M, Parry GC, Stromblad S, Mackman N, Rosenberg S, Mueller BM, Cheresh DA: **Requirement of receptor-bound urokinase-type plasminogen activator for integrin alphavbeta5-directed cell migration**. *J Biol Chem* 1996, **271**(46):29393-29399.

108. Waltz DA, Chapman HA: **Reversible cellular adhesion to vitronectin linked to urokinase receptor occupancy**. *J Biol Chem* 1994, **269**(20):14746-14750.

109. Wei Y, Lukashev M, Simon DI, Bodary SC, Rosenberg S, Doyle MV, Chapman HA: **Regulation of integrin function by the urokinase receptor**. *Science* 1996, **273**(5281):1551-1555.

110. Pfaff M, Du X, Ginsberg MH: **Calpain cleavage of integrin beta cytoplasmic domains**. *FEBS Lett* 1999, **460**(1):17-22.

111. Deryugina EI, Bourdon MA, Reisfeld RA, Strongin A: **Remodeling of collagen matrix by human tumor cells requires activation and cell surface association of matrix metalloproteinase-2**. *Cancer Res* 1998, **58**(16):3743-3750.

112. Chew TL, Wolf WA, Gallagher PJ, Matsumura F, Chisholm RL: **A fluorescent resonant energy transfer-based biosensor reveals transient and regional myosin light chain kinase activation in lamella and cleavage furrows**. *J Cell Biol* 2002, **156**(3):543-553.

113. Dvorak HF, Nagy JA, Feng D, Brown LF, Dvorak AM: **Vascular permeability factor/vascular endothelial growth factor and the significance of microvascular hyperpermeability in angiogenesis**. *Curr Top Microbiol Immunol* 1999, **237**:97-132.

114. Rofstad EK, Tunheim SH, Mathiesen B, Graff BA, Halsor EF, Nilsen K, Galappathi K: **Pulmonary and lymph node metastasis is associated with primary tumor interstitial fluid pressure in human melanoma xenografts**. *Cancer Res* 2002, **62**(3):661-664.

115. Weis S, Cui J, Barnes L, Cheresh D: **Endothelial barrier disruption by VEGF-mediated Src activity potentiates tumor cell extravasation and metastasis**. *J Cell Biol* 2004, **167**(2):223-229.

116. Koop S, MacDonald IC, Luzzi K, Schmidt EE, Morris VL, Grattan M, Khokha R, Chambers AF, Groom AC: **Fate of melanoma cells entering the microcirculation: over 80% survive and extravasate**. *Cancer Res* 1995, **55**(12):2520-2523.

117. Morris VL, Koop S, MacDonald IC, Schmidt EE, Grattan M, Percy D, Chambers AF, Groom AC: **Mammary carcinoma cell lines of high and low metastatic potential differ not in extravasation but in subsequent migration and growth**. *Clin Exp Metastasis* 1994, **12**(6):357-367.

118. Gordan JD, Simon MC: **Hypoxia-inducible factors: central regulators of the tumor phenotype**. *Curr Opin Genet Dev* 2007, **17**(1):71-77.

119. Vaupel P: **The role of hypoxia-induced factors in tumor progression**. *Oncologist* 2004, **9 Suppl 5**:10-17.

120. Cairns RA, Kalliomaki T, Hill RP: **Acute (cyclic) hypoxia enhances spontaneous metastasis of KHT murine tumors**. *Cancer Res* 2001, **61**(24):8903-8908.

121. Postovit LM, Adams MA, Lash GE, Heaton JP, Graham CH: **Oxygen-mediated regulation of tumor cell invasiveness. Involvement of a nitric oxide signaling pathway**. *J Biol Chem* 2002, **277**(38):35730-35737.

122. Rofstad EK, Rasmussen H, Galappathi K, Mathiesen B, Nilsen K, Graff BA: **Hypoxia promotes lymph node metastasis in human melanoma xenografts by up-regulating the urokinase-type plasminogen activator receptor**. *Cancer Res* 2002, **62**(6):1847-1853.

123. Brown JM, Giaccia AJ: **The unique physiology of solid tumors: opportunities (and problems) for cancer therapy**. *Cancer Res* 1998, **58**(7):1408-1416.

124. Esteban MA, Tran MG, Harten SK, Hill P, Castellanos MC, Chandra A, Raval R, O'Brien T S, Maxwell PH: **Regulation of E-cadherin expression by VHL and hypoxia-inducible factor**. *Cancer Res* 2006, **66**(7):3567-3575.

125. Kurrey NK, K A, Bapat SA: **Snail and Slug are major determinants of ovarian cancer invasiveness at the transcription level**. *Gynecol Oncol* 2005, **97**(1):155-165.

126. Luo Y, He DL, Ning L, Shen SL, Li L, Li X: **Hypoxia-inducible factor-1alpha induces the epithelial-mesenchymal transition of human prostatecancer cells**. *Chin Med J (Engl)* 2006, **119**(9):713-718.

127. Zagorska A, Dulak J: **HIF-1: the knowns and unknowns of hypoxia sensing**. *Acta Biochim Pol* 2004, **51**(3):563-585.

128. Tsutsumi S, Yanagawa T, Shimura T, Kuwano H, Raz A: **Autocrine motility factor signaling enhances pancreatic cancer metastasis**. *Clin Cancer Res* 2004, **10**(22):7775-7784.

129. Graham CH, Fitzpatrick TE, McCrae KR: **Hypoxia stimulates urokinase receptor expression through a heme protein-dependent pathway**. *Blood* 1998, **91**(9):3300-3307.

130. Krishnamachary B, Berg-Dixon S, Kelly B, Agani F, Feldser D, Ferreira G, Iyer N, LaRusch J, Pak B, Taghavi P *et al*: **Regulation of colon carcinoma cell invasion by hypoxia-inducible factor 1**. *Cancer Res* 2003, **63**(5):1138-1143.

131. Lee KH, Choi EY, Hyun MS, Kim JR: **Involvement of MAPK pathway in hypoxia-induced up-regulation of urokinase plasminogen activator receptor in a human prostatic cancer cell line, PC3MLN4**. *Exp Mol Med* 2004, **36**(1):57-64.

132. Forsythe JA, Jiang BH, Iyer NV, Agani F, Leung SW, Koos RD, Semenza GL: **Activation of vascular endothelial growth factor gene transcription by hypoxia-inducible factor 1**. *Mol Cell Biol* 1996, **16**(9):4604-4613.

133. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**(5235):467-470.

134. Lewis JD, Izaurralde E: **The role of the cap structure in RNA processing and nuclear export**. *Eur J Biochem* 1997, **247**(2):461-469.

135. David CJ, Manley JL: **Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged**. *Genes Dev* 2010, **24**(21):2343-2364.

136. Ekins R, Chu FW: **Microarrays: their origins and applications**. *Trends Biotechnol* 1999, **17**(6):217-218.

137. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D: **Light-directed, spatially addressable parallel chemical synthesis**. *Science* 1991, **251**(4995):767-773.

138. McGillis DA: **Lithography**. New York: McGraw-Hill; 1983.

139. Southern EM: **Genome mapping: cDNA approaches**. *Curr Opin Genet Dev* 1992, **2**(3):412-416.

140. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale**. *Science* 1997, **278**(5338):680-686.

141. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.

142. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome**. *Science* 2001, **291**(5507):1304-1351.

143. **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**(7011):931-945.

144. Kronick MN: **Creation of the whole human genome microarray**. *Expert Rev Proteomics* 2004, **1**(1):19-28.

145. Norton PA: **Alternative pre-mRNA splicing: factors involved in splice site selection**. *J Cell Sci* 1994, **107 (Pt 1)**:1-7.

146. le Sage C, Agami R: **Immense promises for tiny molecules: uncovering miRNA functions**. *Cell Cycle* 2006, **5**(13):1415-1421.

147. Cai Y, Yu X, Hu S, Yu J: **A brief review on the mechanisms of miRNA regulation**. *Genomics Proteomics Bioinformatics* 2009, **7**(4):147-154.

148. Khraiwesh B, Arif MA, Seumel GI, Ossowski S, Weigel D, Reski R, Frank W: **Transcriptional control of gene expression by microRNAs**. *Cell* 2010, **140**(1):111-122.

149. Lin J, Xie Z, Zhu H, Qian J: **Understanding protein phosphorylation on a systems level**. *Brief Funct Genomics* 2010, **9**(1):32-42.

150. Rocca-Serra P, Brazma A, Parkinson H, Sarkans U, Shojatalab M, Contrino S, Vilo J, Abeygunawardena N, Mukherjee G, Holloway E *et al*: **ArrayExpress: a public database of gene expression data at EBI**. *C R Biol* 2003, **326**(10-11):1075-1078.

151. Edgar R, Barrett T: **NCBI GEO standards and services for microarray data**. *Nat Biotechnol* 2006, **24**(12):1471-1472.

152. Lin WC: **Les puces à ADN sur lames de verre: principes et méthodes de confection, d'application expérimentale et d'analyse des données**. Institut Curie; 2004.

153. Zghidi W, Merendino L, Cottet A, Mache R, Lerbs-Mache S: **Nucleus-encoded plastid sigma factor SIG3 transcribes specifically the psbN gene in plastids**. *Nucleic Acids Res* 2007, **35**(2):455-464.

154. Magbanua ZV, Ozkan S, Bartlett BD, Chouvarine P, Saski CA, Liston A, Cronn RC, Nelson CD, Peterson DG: **Adventures in the enormous: a 1.8 million clone BAC library for the 21.7 Gb genome of loblolly pine**. *PLoS One* 2011, **6**(1):e16214.

155. Schroder C, Jacob A, Tonack S, Radon TP, Sill M, Zucknick M, Ruffer S, Costello E, Neoptolemos JP, Crnogorac-Jurcevic T *et al*: **Dual-color proteomic profiling of complex samples with a microarray of 810 cancer-related antibodies**. *Mol Cell Proteomics* 2010, **9**(6):1271-1280.

156. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: **High density synthetic oligonucleotide arrays**. *Nat Genet* 1999, **21**(1 Suppl):20-24.

157. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR *et al*: **Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer**. *Nat Biotechnol* 2001, **19**(4):342-347.

158. Iida K, Nishimura I: **Gene expression profiling by DNA microarray technology**. *Crit Rev Oral Biol Med* 2002, **13**(1):35-50.

159. Midorikawa Y, Makuuchi M, Tang W, Aburatani H: **Microarray-based analysis for hepatocellular carcinoma: from gene expression profiling to new challenges**. *World J Gastroenterol* 2007, **13**(10):1487-1492.

160. Santos GC, Zielenska M, Prasad M, Squire JA: **Chromosome 6p amplification and cancer progression**. *J Clin Pathol* 2007, **60**(1):1-7.

161. Maskos U, Southern EM: **Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ**. *Nucleic Acids Res* 1992, **20**(7):1679-1684.

162. Wu J, Smith LT, Plass C, Huang TH: **ChIP-chip comes of age for genome-wide functional analysis**. *Cancer Res* 2006, **66**(14):6899-6902.

163. Lai E: **Application of SNP technologies in medicine: lessons learned and future challenges**. *Genome Res* 2001, **11**(6):927-929.

164. Kechris K, Yang YH, Yeh RF: **Prediction of alternatively skipped exons and splicing enhancers from exon junction arrays**. *BMC Genomics* 2008, **9**:551.

165. Laajala E, Aittokallio T, Lahesmaa R, Elo LL: **Probe-level estimation improves the detection of differential splicing in Affymetrix exon array studies**. *Genome Biol* 2009, **10**(7):R77.

166. Cin H, Meyer C, Herr R, Janzarik WG, Lambert S, Jones DT, Jacob K, Benner A, Witt H, Remke M *et al*: **Oncogenic FAM131B-BRAF fusion resulting from 7q34 deletion comprises an alternative mechanism of MAPK pathway activation in pilocytic astrocytoma**. *Acta Neuropathol* 2011, **121**(6):763-774.

167. Affymetrix: **Affymetrix Microarray Suite User Guide version 5.0**. Santa Clara: Affymetrix Manual; 2001.

168. Leung YF, Cavalieri D: **Fundamentals of cDNA microarray data analysis**. *Trends Genet* 2003, **19**(11):649-659.

169. Gautier L, Moller M, Friis-Hansen L, Knudsen S: **Alternative mapping of probes to genes for Affymetrix chips**. *BMC Bioinformatics* 2004, **5**:111.

170. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H *et al*: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data**. *Nucleic Acids Res* 2005, **33**(20):e175.

171. Liu H, Zeeberg BR, Qu G, Koru AG, Ferrucci A, Kahn A, Ryan MC, Nuhanovic A, Munson PJ, Reinhold WC *et al*: **AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets**. *Bioinformatics* 2007, **23**(18):2385-2390.

172. Hochreiter S, Clevert DA, Obermayer K: **A new summarization method for Affymetrix probe level data**. *Bioinformatics* 2006, **22**(8):943-949.

173. Yang YH, Buckley MJ, Dudoit S, Speed TP: **Comparison of methods for image analysis on cDNA microarray data**. *Journal of Computational and Graphical Statistics* 2002, **11**(1):108-136.

174. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.

175. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**(2):249-264.

176. Freudenberg J, Boriss H, Hasenclever D: **Comparison of preprocessing procedures for oligo-nucleotide micro-arrays by parametric bootstrap simulation of spike-in experiments**. *Methods Inf Med* 2004, **43**(5):434-438.

177. Wu Z, Irizarry R, Gentleman R, Murillo F, Spencer F: **A model-based background adjustment for oligonucleotide expression arrays**. *Journal of the American Statistical Association* 2004, **99**:909-917.

178. Draghici S: **Statistical intelligence: effective analysis of high-density microarray data**. *Drug Discov Today* 2002, **7**(11 Suppl):S55-63.

179. Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM: **Microarray expression profiling identifies genes with altered expression in HDL-deficient mice**. *Genome Res* 2000, **10**(12):2022-2029.

180. Student: **The Probable Error of a Mean**. *Biometrika* 1908, **6**:1-25.

181. Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes**. *Bioinformatics* 2001, **17**(6):509-519.

182. Berger F, De Hertogh B, Pierre M, Gaigneaux A, Depiereux E: **The "Window t test": a simple and powerful approach to detect differentially expressed genes in microarray datasets**. *Central European Journal of Biology* 2008, **3**(3):327-344.

183. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

184. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Res* 1999, **27**(1):29-34.

185. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC: **A global test for groups of genes: testing association with a clinical outcome**. *Bioinformatics* 2004, **20**(1):93-99.

186. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA *et al*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science* 1999, **286**(5439):531-537.

187. Goeman JJ, Oosting J, Cleton-Jansen AM, Anninga JK, van Houwelingen HC: **Testing association of a pathway with survival using gene expression data**. *Bioinformatics* 2005, **21**(9):1950-1957.

188. Mansmann U, Meister R: **Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach**. *Methods Inf Med* 2005, **44**(3):449-453.

189. Berger F, De Meulder B, Gaigneaux A, Depiereux S, Bareke E, Pierre M, De Hertogh B, Delorenzi M, Depiereux E: **Functional analysis: evaluation of response intensities--tailoring ANOVA for lists of expression subsets**. *BMC Bioinformatics* 2010, **11**:510.

190. Opgen-Rhein R, Strimmer K: **Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach**. *Stat Appl Genet Mol Biol* 2007, **6**:Article9.

191. Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS: **Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset**. *Genome Biol* 2005, **6**(2):R16.

192. De Hertogh B, De Meulder B, Berger F, Pierre M, Bareke E, Gaigneaux A, Depiereux E: **A benchmark for statistical microarray data analysis that preserves actual biological and technical variance**. *BMC Bioinformatics* 2010, **11**:17.

193. Dunn OJ: **Multiple Comparisons Among Means**. *Journal of the American Statistical Association* 1961, **56**:52-64.

194. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society* 1995, **57**:289-300.

195. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biol* 2003, **4**(5):P3.

196. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc* 2009, **4**(1):44-57.

197. Bareke E, Pierre M, Gaigneaux A, De Meulder B, Depiereux S, Berger F, Habra N, Depiereux E: **PathEx: a novel multi factors based datasets selector web tool**. *BMC Bioinformatics* 2010, **11**:528.

198. Faure AJ, Seoighe C, Mulder NJ: **Investigating the effect of paralogs on microarray gene-set analysis**. *BMC Bioinformatics* 2011, **12**:29.

199. Gur-Dedeoglu B, Konu O, Kir S, Ozturk AR, Bozkurt B, Ergul G, Yulug IG: **A resampling-based meta-analysis for detection of differential gene expression in breast cancer**. *BMC Cancer* 2008, **8**:396.

200. Ma S, Huang J: **Regularized gene selection in cancer microarray meta-analysis**. *BMC Bioinformatics* 2009, **10**:1.

201. Ochsner SA, Steffen DL, Hilsenbeck SG, Chen ES, Watkins C, McKenna NJ: **GEMS (Gene Expression MetaSignatures), a Web resource for querying meta-analysis of expression microarray datasets: 17beta-estradiol in MCF-7 cells**. *Cancer Res* 2009, **69**(1):23-26.

202. Sims AH, Smethurst GJ, Hey Y, Okoniewski MJ, Pepper SD, Howell A, Miller CJ, Clarke RB: **The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis**. *BMC Med Genomics* 2008, **1**:42.

203. Hunter JE, Schmidt FL: **Methods of Meta-Analysis: Correcting Error and Bias in Research Findings**. Newbury Park, California: SAGE Publications; 1990.

204. Rosenthal, Robert: **The "File Drawer Problem" and the Tolerance for Null Results**. *Psychological Bulletin* 1979, **86**(3):638-641.

205. Roseman M, Milette K, Bero LA, Coyne JC, Lexchin J, Turner EH, Thombs BD: **Reporting of conflicts of interest in meta-analyses of trials of pharmacological treatments**. *Jama* 2011, **305**(10):1008-1017.

206. Provenzani A, Fronza R, Loreni F, Pascale A, Amadio M, Quattrone A: **Global alterations in mRNA polysomal recruitment in a cell model of colorectal cancer progression to metastasis**. *Carcinogenesis* 2006, **27**(7):1323-1333.

207. Pierre M, DeHertogh B, Gaigneaux A, DeMeulder B, Berger F, Bareke E, Michiels C, Depiereux E: **Meta-analysis of archived DNA microarrays identifies genes regulated by hypoxia and involved in a metastatic phenotype in cancer cells**. *BMC Cancer* 2010, **10**:176.

208. Pierre M, DeHertogh B, DeMeulder B, Bareke E, Depiereux S, Michiels C, Depiereux E: **Enhanced Meta-analysis Highlights Genes Involved in Metastasis from Several Microarray Datasets**. *Journal of Proteomics and Bioinformatics* 2011, **4**(2):036-043.

209. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY *et al*: **The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements**. *Nat Biotechnol* 2006, **24**(9):1151-1161.

210. **MAQC-II: analyze that!** *Nat Biotechnol* 2010, **28**(8):761.

211. Luo J, Schumacher M, Scherer A, Sanoudou D, Megherbi D, Davison T, Shi T, Tong W, Shi L, Hong H *et al*: **A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data**. *Pharmacogenomics J* 2010, **10**(4):278-291.

212. Cahan P, Rovegno F, Mooney D, Newman JC, St Laurent G, 3rd, McCaffrey TA: **Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization**. *Gene* 2007, **401**(1-2):12-18.

213. Larsson O, Sandberg R: **Lack of correct data format and comparability limits future integrative microarray research**. *Nat Biotechnol* 2006, **24**(11):1322-1323.

214. Wennmalm K, Wahlestedt C, Larsson O: **The expression signature of in vitro senescence resembles mouse but not human aging**. *Genome Biol* 2005, **6**(13):R109.

215. Jenner RG, Young RA: **Insights into host responses against pathogens from transcriptional profiling**. *Nat Rev Microbiol* 2005, **3**(4):281-294.

216. French SW, Oliva J, French BA, Li J, Bardag-Gorce F: **Alcohol, nutrition and liver cancer: role of Toll-like receptor signaling**. *World J Gastroenterol* 2010, **16**(11):1344-1348.

217. Zheng SL, Augustsson-Balter K, Chang B, Hedelin M, Li L, Adami HO, Bensen J, Li G, Johnasson JE, Turner AR *et al*: **Sequence variants of toll-like receptor 4 are associated with prostate cancer risk: results from the CAncer Prostate in Sweden Study**. *Cancer Res* 2004, **64**(8):2918-2922.

218. Chang YJ, Wu MS, Lin JT, Sheu BS, Muta T, Inoue H, Chen CC: **Induction of cyclooxygenase-2 overexpression in human gastric epithelial cells by Helicobacter pylori involves TLR2/TLR9 and c-Src-dependent nuclear factor-kappaB activation**. *Mol Pharmacol* 2004, **66**(6):1465-1477.

219. Huang B, Zhao J, Li H, He KL, Chen Y, Chen SH, Mayer L, Unkeless JC, Xiong H: **Toll-like receptors on tumor cells facilitate evasion of immune surveillance**. *Cancer Res* 2005, **65**(12):5009-5014.

220. May RC, Machesky LM: **Phagocytosis and the actin cytoskeleton**. *J Cell Sci* 2001, **114**(Pt 6):1061-1077.